STIFTUNG DATENSCHUTZ

PRAXISLEITFADEN ZUM ANONYMISIEREN PERSONENBEZOGENER DATEN

Anforderungen, Einsatzklassen und Vorgehensmodell

VON

Prof. Dr. Rolf Schwartmann, Andreas Jaspers, Dr. Niels Lepperhoff, Steffen Weiß LL.M., Prof. Dr. Michael Meier

Praxisleitfaden zum Anonymisieren personenbezogener Daten

Anforderungen, Einsatzklassen und Vorgehensmodell

erstellt und vorgelegt im Auftrag der Stiftung Datenschutz

im Dezember 2022

von

Professor Dr. Rolf Schwartmann
Kölner Forschungsstelle für Medienrecht, Technische Hochschule Köln,
Gesellschaft für Datenschutz und Datensicherheit (GDD) e.V.,

Rechtsanwalt Andreas Jaspers
Gesellschaft für Datenschutz und Datensicherheit (GDD) e.V.,
DSZ Datenschutz Zertifizierungsgesellschaft mbH,

Dr. Niels Lepperhoff
DSZ Datenschutz Zertifizierungsgesellschaft mbH,
XAMIT Bewertungsgesellschaft mbH,

Rechtsanwalt Steffen Weiß, LL.M.
Gesellschaft für Datenschutz und Datensicherheit (GDD) e.V.

unter Mitwirkung von

Professor Dr. Michael Meier*
Universität Bonn, Fraunhofer FKIE,
Gesellschaft für Datenschutz und Datensicherheit (GDD) e.V.

*Mit Beteiligung von Saffija Kasem-Madani, Markus Krämer, Daniel Meyer (alle Universität Bonn)

Inhaltsverzeichnis

1. Einleitung

Daten stehen im Mittelpunkt des digitalen Wandels. Der Austausch personenbezogener Daten zwischen öffentlichen und privaten Stellen nimmt stetig zu. Die europäische Datenstrategie soll der EU zu einer Führungsposition im internationalen Wettbewerb um datengetriebene Geschäftsmodelle verhelfen.

Das Ziel der DS-GVO ist es, die Verarbeitung personenbezogener Daten in den Dienst der Menschheit zu stellen. Darunter fallen Zwecke des Gemeinwohls ebenso wie unternehmerische Belange. Die **DS-GVO** will Datenverarbeitung im Einklang mit den praktischen und wirtschaftlichen Erfordernissen **ermöglichen, nicht verhindern**. Der Datenschutz ist insofern, um ein Bild des vernetzten Fahrens zu benutzen, eher Spurassistent als Bremse. Die DS-GVO ist für neue datengetriebene wirtschaftliche und technische Entwicklungen offen. Sie steht neuen Vertragskonstruktionen und damit einhergehenden Datenverarbeitungen auch unter der Überschrift „Zahlen mit Daten“ nicht entgegen. Anders sind auf Datenmaximierung angelegte Dienste wie Soziale Netzwerke rechtlich nicht zu fassen. Gestattet sind nach entsprechender Abwägung zudem Weiterverarbeitungen von Daten zu interessengerechten und kompatiblen neuen Zwecken.

Das Datenschutzrecht ermöglicht dabei keineswegs nur die Verarbeitung anonymer Daten, die ganz ohne Personenbezug auskommen. Namentlich **Verschlüsselung und Pseudonymisierung** von Personendaten dienen als **„Ermöglichungswerkzeuge“** zur technischen Absicherung des Schutzes bei der Verarbeitung auch großer Datenmengen. Die Datenethikkommission der Bundesregierung hat dem Thema des Datenzugangs sowohl für personenbezogene Daten als auch für nicht personenbezogene Daten im Abschnitt E 4. und E 5. wichtigen Raum eingeräumt.[1]

Politisch ist insbesondere die datenschutzkonforme Auswertung von Gesundheitsdaten von Krankheiten erwünscht. Die Europäische Kommission hat im Frühjahr 2022 den Entwurf einer Verordnung zur Schaffung eines europäischen Raums für Gesundheitsdaten vorgestellt. Sie soll Einzelpersonen die Kontrolle über ihre Gesundheitsdaten ermöglichen und zugleich die Nutzung von Gesundheitsdaten für bessere medizinische Versorgung und Forschung eröffnen. Die EU soll das Potenzial von Austausch, Nutzung und Weiterverwendung von Gesundheitsdaten ausschöpfen. Auch die Datenethikkommission der Bundesregierung macht sich hierfür unter 3.5.2 des Abschlussgutachtens stark.

Der Koalitionsvertrag der deutschen Ampelregierung hat das aufgegriffen. Die Fortentwicklung der **Digitalisierung des Gesundheitswesens** ist vereinbart. Zu Recht, denn spätestens die Pandemiebekämpfung hat uns vor Augen geführt, wie wichtig es ist, einen ausgewogenen Rahmen für die Verarbeitung von Gesundheitsdaten zu schaffen, der die Belange des Gesundheitsschutzes in ein angemessenes Verhältnis zum Schutz der Privatheit setzt.

[1] Abschlussgutachten der Datenethikkommission (2019).

Darüber hinaus bestehen weitere praktische Anwendungsfälle der Anonymisierung, von der Verwendung von digitalen Straßenkarten, bis hin zu aggregierten Nutzerstatistiken im Online-Bereich oder im Rahmen vertraglicher Kundenbeziehungen.

Hier setzen die Grundregeln dieser Untersuchung im Auftrag der Stiftung Datenschutz über die **Anonymisierung** an. Der Gegenstand der vorliegenden Arbeit zur Anonymisierung soll die **Basis für flächendeckende Anwendungen** in der Praxis sein. Anonymisierung von Daten macht den Rückschluss auf eine Person unmöglich und nimmt sie vom Anwendungsbereich der DS-GVO aus. Das ist dann gewünscht und erforderlich, wenn anonyme Daten den Zweck nach der Verarbeitung erfüllen. Sofern betroffene Personen etwa zur eigenen Gesundheitsvorsorge den Rückschluss auf ihre Daten erhalten möchten, sieht die DS-GVO deren Pseudonymisierung vor. Sie sorgt dafür, dass Daten durch Verschlüsselung gegen Missbrauch geschützt werden. Diesbezüglich verweist die Studie auf den von der Fokusgruppe Datenschutz des Bundesinnenministeriums im Rahmen des Digital-Gipfels 2019 erarbeiteten **„Entwurf für einen Code of Conduct zum Einsatz DS-GVO-konformer Pseudonymisierung“**.

Dreh- und Angelpunkt des gesetzlichen Anwendungsbereichs ist die Frage des Personenbezugs. Wird dieser entfernt, können die gesetzlichen Vorgaben keine Wirksamkeit beanspruchen. Nach den Erwägungen der DS-GVO soll der Datenschutz für alle Informationen gelten, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Liegen diese Merkmale nicht vor, ist ein Datum anonym. Für die Weiterverwendung personenbezogener Daten ist von Bedeutung, dass das Gesetz auch die Anonymisierung von personenbezogenen Daten als möglich erachtet (vgl. Erwägungsgrund 26). Nach der Umwandlung oder Veränderung der personenbezogenen Daten kann der Betroffene nicht oder nicht mehr identifiziert werden.

Der Wunsch von Verantwortlichen per Anonymisierung den Anwendungsbereich der DS-GVO zu verlassen, um Daten einfacher durch sich oder durch Dritte nutzbar zu machen, ist nachvollziehbar. Umso wichtiger ist es aber, die Grenze des rechtlich zulässigen, zunächst was deren Grundregeln anbelangt, sauber zu ziehen. Nur so kann Rechtsklarheit und Rechtssicherheit unter Wahrung der Erfordernisse der DS-GVO geschaffen werden.

Dieser Praxisleitfaden befasst sich mit Hinweisen zur Anonymisierung von personenbezogenen Daten.[2] Hierzu wird der **Begriff der Anonymisierung** und dessen Ausprägungen im bestehenden rechtlichen Kontext **eingeordnet**. Dabei muss eine Abgrenzung von anderen Verarbeitungsvorgängen erfolgen, namentlich zur Pseudonymisierung. Nach der begrifflichen Einordnung werden gängige **Verfahren und Techniken einer Anonymisierung** allgemein beschrieben. Um den Praxisbezug zu wahren schließen sich hieran Einsatzklassen einer Anonymisierung an. Hierbei werden Anwendern **Einsatzszenarien und -beispiele** aufgezeigt, in denen eine Anonymisierung erfolgen kann. Ein gesondertes Kapitel wird sich mit dem rechtlichen Umfeld der Anonymisierung

[2] Zu den Anforderungen an die Pseudonymisierung Schwartmann/Weiß, Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen - ein Arbeitspapier der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2018.

befassen und den dabei bestehenden Anforderungen, seien es besondere Prüf-, Dokumentations- oder Transparenzpflichten. Um insbesondere **kleinere und mittelständische Unternehmen zu unterstützen**, wird ein Vorgehensmodell zur Verfügung gestellt, um den Vorgang der Anonymisierung schrittweise und strukturiert zu vollziehen.

Die nachfolgende Darstellung dient als allgemeine Orientierungshilfe bei der Anonymisierung personenbezogener Daten. Sie kann und soll nicht dazu dienen, abschließende Vorgaben für ein solches Verfahren zu formulieren. Darüber hinaus können sich aus anderen Gesetzen Vorgaben für nicht personenbezogene Daten ergeben, die entsprechend auf einen anonymisierten Datensatz Anwendung finden, aktuell etwa bereits im Data Governance Act.

2. Der Praxisleitfaden im Überblick

2.1 Rahmenbedingungen der Anonymisierung

Die Anonymisierung personenbezogener Daten ist ein vielschichtiges Verfahren. Sie bedarf zunächst eines grundlegenden Verständnisses zentraler **Begrifflichkeiten** mit Blick auf deren Rahmenbedingungen**.** Dies beginnt beim grundlegenden Verständnis von Daten, die in strukturierter sowie unstrukturierter Form vorliegen können.

- **Strukturierte Daten** können als Schlüssel-Wert-Paar verstanden werden, während **unstrukturierte Daten** alle Daten sind, die nicht einer Schlüssel-Wert-Paar-Darstellung entsprechen.
- **Personenbezogene Daten**, als eine weitere Ausprägung von Daten, sind durch die DS-GVO definiert und weisen durch den gesetzgeberischen Hinweis auf eine „Identifizierbarkeit“ einen weiten Anwendungsbereich auf.
- Während die **Anonymisierung** zum Ziel hat, einen Personenbezug zu entfernen, ist eine Re-Identifizierung Betroffener im Rahmen der **Pseudonymisierung** weiterhin möglich. Allerdings müssen die zur Re-Identifizierung verwendbaren Daten hierbei gesondert aufbewahrt und durch technisch-organisatorische Maßnahmen geschützt sein.
- Der **Übergang** zwischen Anonymisierung und Pseudonymisierung kann mitunter **fließend** sein. Insbesondere wenn eine Verarbeitung in einem anderen Kontext stattfindet und zusätzlichen Informationen vorhanden sind, die zur Identifizierung einer betroffenen Person beitragen.

2.2 Begriffe der Anonymisierung unter Einbeziehung von Wertungen

Es existieren verschiedene Ansätze **einer Anonymisierung**.

- Ist der Personenbezug praktisch für jedermann unmöglich, spricht man von **absoluter Anonymisierung**.
- Die **faktische bzw. relative Anonymisierung** zeichnet sich dadurch aus, dass die Re-Identifizierbarkeit der betroffenen Person nicht gänzlich ausgeschlossen ist. Allerdings scheidet eine Re-Identifizierung der betroffenen Person aufgrund der **Unverhältnismäßigkeit ihres Aufwandes** aus.

Die DS-GVO eröffnet in ihren Erwägungsgründen (ErwG) die Möglichkeit, Verhältnismäßigkeitsaspekte in die Frage einer erfolgreichen Anonymisierung mit einzubeziehen. Die **höchstrichterliche Rechtsprechung** hält es nicht für erforderlich, dass sich sämtliche, zur Identifizierung erforderlichen Informationen in der Hand eines einzigen Verantwortlichen befinden. Ausreichend ist vielmehr, wenn der Verantwortliche über einen Dritten die betroffene Person bestimmen lässt. Die Grenze der Identifizierbarkeit liegt wiederum in der Unmöglichkeit, der Unverhältnismäßigkeit oder im Rechtsverstoß.

2.3 Sonderfall: Künstliche Intelligenz und verwandte Techniken

Die Diskussion der Anonymisierung im Rahmen der **Künstlichen Intelligenz (KI)** setzt ein Verständnis von Hintergründen dieser neueren Technologie voraus. Grundsätzlich ist KI nichts anderes als ein solches Computerprogramm. Hierbei existieren eine Vielzahl von unterschiedlichen Arten von **Algorithmen und Vorgehensweisen**, die jeweils anders funktionieren und unterschiedliche Einsatzgebiete haben. Zu der heute in der öffentlichen Wahrnehmung domminierenden Algorithmen-Klasse zählt die Mustererkennung mittels maschinellen Lernens. Die Algorithmen zur Mustererkennung enthalten unspezifische Regeln, die erst in einer „Trainingsphase“ angepasst werden. Auch Algorithmen und ein hieraus ermitteltes Ergebnis unterliegen den Abgrenzungsschwierigkeiten zwischen personenbezogenen und anonymen Daten.

2.4 Europarechtlicher Kontext: DS-GVO und „neue Datenakte“

2.4.1 Neue Datenakte

Die Anonymisierung personenbezogener Daten bewegt sich nicht nur in einem technischen sondern auch in einem rechtlichen Umfeld. Die **europäische Datenstrategie** der Kommission zielt auf neuere Technologien, z.B. über den Entwurf einer KI-Verordnung, sowie das Datenteilen mittels Data Governance Act und dem Entwurf für einen Data Act ab. Hierbei sollen neue **Datenräume** für bestimmte Sektoren entstehen und gesetzliche Anreize wie Verpflichtungen für die Weitergabe personenbezogener wie nicht personenbezogener Daten geschaffen werden. Die in den neuen Rechtsakten vorgesehenen Mechanismen für das Datenteilen warten zumeist mit der Möglichkeit **zur Implementierung von Schutzmaßnahmen** zugunsten Betroffener auf. Hierzu gehören auch die **Pseudonymisierung** und **Anonymisierung** personenbezogener Daten. Die Anonymisierung von Daten wird also mit den neuen Datenakten signifikant an Bedeutung gewinnen. Die DS-GVO bleibt der Standard, wenn es um die Verarbeitung personenbezogener Daten geht, der zusätzlich zu spezifischen Vorgaben der neuen Rechtsakte einzuhalten ist.

2.4.2 DS-GVO

Als weiterhin bestehender Standard für die Anonymisierung fordert die **DS-GVO** eine **Rechtsgrundlage** für das Anonymisieren von personenbezogenen Daten. Da die Zwecke einer initialen Erhebung personenbezogener Daten und derer der Anonymisierung

regelmäßig voneinander abweichen, muss geprüft werden, ob diese **Zwecke kompatibel sind**. Von einer solchen Kompatibilität ist regelmäßig auszugehen, sollten keine besonderen Kategorien personenbezogener Daten der Anonymisierung zugrunde liegen.

Grundsätzlich wird durch die Anonymisierung der Anwendungsbereich der DS-GVO verlassen. In manchen Fällen bestehen jedoch Unsicherheiten, ob von einer erfolgreichen Anonymisierung nach den gesetzlichen Standards ausgegangen werden kann. Verantwortliche sind nicht daran gehindert, die Anonymisierung **als technisch-organisatorische Maßnahme** zu begreifen, die auf personenbezogene Daten angewendet wird. Hierdurch bewegt sich ein Verantwortlicher zwar weiterhin im gesetzlichen Anwendungsbereich, er kann sich jedoch die getätigten Schutzmaßnahmen bspw. im Rahmen einer Interessensabwägung zugutehalten lassen.

Die **Anforderungen** an die Anonymisierung personenbezogener Daten können grundsätzlich in einen **rechtlichen und technischen Teil** kategorisiert werden. Da jedoch ein Interesse Dritter an anonymisierten Daten von besonderer Bedeutung für die Wahl der technischen Mittel einer Anonymisierung darstellt, wird das Beschreiben eines **Angreifermodells** durch den Praxisleitfaden als weitere Anforderung aufgeführt.

Aus rechtlicher Sicht besteht **eine Prüfpflicht**, ob sich die aus einer Anonymisierung erzeugten Daten auf eine identifizierte oder identifizierbare Person beziehen. Im Rahmen dieser Prüfung sind **alle Mittel** zu berücksichtigen, die vernünftigerweise entweder von dem Verantwortlichen oder einem Dritten eingesetzt werden könnten, um die betroffene Person zu identifizieren. Solche Mittel können bspw. für den Verantwortlichen verfügbare Informationen sein, oder solche, die er sich beschaffen kann. Auch mögliche Verknüpfungen von Daten sind in die Prüfung mit einzubeziehen. Gerade mit Blick auf die **europäische Datenstrategie** sind öffentlich-zugängliche Datenräume mit personenbezogenen, pseudonymisierten oder anonymisierten Daten vermehrt zu erwarten. Eine Vielzahl von Datenquellen kann die Wahrscheinlichkeit einer Re-Identifizierung erhöhen. Irrelevant ist, ob der Verantwortliche oder ein Empfänger Daten der Person identifizieren möchte oder nicht. Es reicht eine objektive Identifizierbarkeit aus.

Aus Sicht der DS-GVO können, mangels gesetzlicher Präzisierung, verschiedene Anonymisierungstechniken eingesetzt werden. Entscheidend ist, dass nach Prüfung der oben aufgeführten Faktoren eine Re-Identifizierung von Betroffenen praktisch nicht durchführbar ist. D.h. erfordert sie einen **unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft**, kann grundsätzlich von einer wirksamen Anonymisierung ausgegangen werden.

2.5 „Angreifermodell“ als Test

Ein **„Angreifermodell”** beschreibt eine Methode, um zu prüfen, ob ein Datensatz anonym oder personenbezogen ist. Aus der Perspektive eines Angreifers wird getestet, ob eine Re-Identifikation möglich ist. Welche Kenntnisse und Fähigkeiten dem Angreifer unterstellt werden, hängt vom Verwendungskontext der Daten ab, z.B. ob anonymisierte Daten veröffentlicht, lediglich intern genutzt oder an bestimmte Empfänger weitergegeben werden. Es empfiehlt sich, beim Angreifermodell nicht nur zielgerichtete

Angriffe zu berücksichtigen, sondern auch Konstellationen, in denen eine Re-Identifizierung durch den Angreifer eigentlich ungewollt ist bzw. zufällig von statten gehen könnte. Es bestehen verschiedene Umsetzungsmöglichkeiten, um einen Angriff mit dem Ziel einer Re-Identifizierung durchzuführen. Hierzu zählen das **Herausgreifen/Aussondern („singling out“) einer Person**, die **Verknüpfung von Datensätzen (“record linkage”)** sowie das Ableiten von Merkmalen einer Person von anderen im Datenbestand vorhandener Merkmale **(Inferenz)**.

2.6 Anonymisierungsmethoden

Aus **technischer Sicht** stehen verschiedene Anonymisierungsmethoden zur Verfügung, die in solche der **Generalisierung** und solcher der **Randomisierung** eingeteilt werden können. Zu den Verfahren der Randomisierung zählen bspw. die **stochastische Überlagerung**, die **Vertauschung von Werten** in einem Datensatz sowie **Differential Privacy**. Zu den Verfahren der Generalisierung zählen mitunter die **Aggregation und k-Anonymität**, **l -Diversität** und **t-Closeness** sowie das Arbeiten mit **synthetischen Daten**. Bei synthetischen Daten handelt es sich um durch Berechnungsverfahren erzeugte Daten, ohne die Identität eines Betroffenen zu offenbaren. Welche Verfahren auf welche Daten anwendbar sind und welche Risiken hinsichtlich des Angreifermodells bestehen, wird im Praxisleitfaden über eine **Bewertungsmatrix** veranschaulicht. Von vornherein **ungeeignete technische Methoden** einer Anonymisierung bestehen. Dies bezieht sich insbesondere auf solche Verfahren, die **Hashfunktionen** zugrunde liegen.

2.7 Szenarien der Datenweitergabe

Vielfach werden anonymisierte Daten an **Dritte weitergegeben**. Ebenso haben Dienstleister ein Interesse an einer Verwendung anonymisierter Daten für eigene Zwecke. Ein Empfänger solcher Daten wird zu überprüfen haben, ob die Daten für ihn unter Berücksichtigung der bei ihm verfügbaren Mittel und der Wahrscheinlichkeit ihres Einsatzes anonym sind. Vertragliche Verbote einer Re-Identifizierung sind im Rahmen dieser objektiven Prüfung kein Kriterium, um eine solche Re-Identifizierung per se auszuschließen. Die Anonymisierung als Verarbeitung personenbezogener Daten kann bei der Einbeziehung Dritter dazu führen, dass mehrere **Verantwortliche gemeinsam für diese Verarbeitung verantwortlich sind**. In einem solchen Fall müssen die Rollen und Verantwortlichkeiten im Rahmen der Anonymisierung in einer Vereinbarung klar beschrieben werden. Das **Delegieren einer Anonymisierung** an einen Auftragsverarbeiter macht eine Anonymisierung angreifbar, da der Verantwortliche jederzeit Weisungen gegenüber dem Dienstleister aussprechen kann, um hierbei eine Offenlegung der verwendeten Technik zu erreichen. Die Folge wäre, dass eine weitere Stelle über das Wissen der Anonymisierungstechnik verfügt, was sich ein Angreifer zunutze machen könnte.

Anonymisiert ein Dienstleister personenbezogene Daten seines Auftraggebers **für eigene Zwecke** schwingt er sich zu einem Verantwortlichen für die Datenverarbeitung auf. Die Folge ist, dass der Verantwortliche eine Rechtsgrundlage für die Übermittlung benötigt und der Dienstleister zur selben Zeit eine Rechtsgrundlage für die Verarbeitung

personenbezogener Daten benötigt. Im Übrigen muss die Zweckänderung mit dem ursprünglichen Zweck **kompatibel** sein.

2.8 Vier Einsatzklassen

Praktische Szenarien einer Anonymisierung personenbezogener Daten lassen sich in **vier Einsatzklassen** unterteilen.

2.8.1 Einsatzklasse 1: Anonymisierung als Löschung

Bei der **Anonymisierung als Löschung** geht es darum, einen Personenbezug in einem Datensatz zu entfernen, um Daten bzw. Eigenschaften in einem Datensatz weiterhin verwenden zu können. Der Praxisleitfaden benennt hierbei drei Beispiele: So das Behalten von Eckdaten von Bewerbungen, die Qualitätsanalyse im Bereich Customer Support sowie die Erstellung von Webseitenstatistiken.

2.8.2 Einsatzklasse 2: Weitergabe anonymisierter Daten

Bei der **Weitergabe anonymisierter Daten**, als weitere Einsatzklasse, werden Szenarien des „Gehaltsbenchmarking“, der Weitergabe von Verkaufszahlen nach Produktkategorie sowie der Abgleich geleakter Zugangsdaten ebenso beschrieben, wie die Analyse eines Kraftstoffverbrauches von Fahrzeugen.

2.8.3 Einsatzklasse 3: Anonymisierung beim Training von Algorithmen

Die dritte Einsatzklasse widmet sich neuen Technologien in Gestalt der **Anonymisierung von Trainingsdaten**. Datenschutzrechtlich stellt sich das Training von Algorithmen aus mehreren Perspektiven als problematisch dar. Grundsätzlich benötigt das präzise Training eines Modells eine breite Datenbasis, um genügend Informationen bereitzustellen. Andernfalls besteht die Gefahr, dass ein zu grobes Modell erstellt wird.

Im Rahmen des Datenschutzes oder des Schutzes von Firmengeheimnissen ist eine Zusammenführung großer Datenmengen jedoch oftmals problematisch, da diese im Unternehmen, welches das Erstellen von Modellen Künstlicher Intelligenz anbietet, bekannt werden und somit in fremde Hände gelangen. An dieser Stelle bietet sich beispielsweise **Federated Learning** an, das im Rahmen des Trainings von Algorithmen zur Anwendung gelangen kann. Im Grundsatz bedeutet **Federated Learning**, dass die Daten sowohl zum Training genutzt werden können als auch beim jeweiligen Dateneigentümer verbleiben. Der Diensteanbieter erstellt zunächst ein initiales Modell, welches er an seine Partner weitergibt. Unter Verwendung seiner jeweiligen Daten testet nun jeder Partner das erhaltene Modell und teilt dem Diensteanbieter mit, wie die Parameter verändert werden sollen, um das Modell zu verbessern. Aus allen erhaltenen Rückmeldungen errechnet der Diensteanbieter nun ein Gesamtupdate und wendet es auf das bisherige Modell an, welches nun erneut verteilt wird.

Um dem Problem der Aufdeckung personenbezogener Daten durch Beobachtung von Veränderungen der Auswertung von Datensätzen im Bereich Federated Learning zu lösen, bietet sich **Differential Privacy** an, d.h. die Vermeidung einer Übermittlung identifizierender Merkmale zu einer Person durch gezielte Übermittlung von Informationen aus einer Datenbank. Als weiteres Beispiel der Einsatzklasse des Trainierens von Algorithmen benennt der Praxisleitfaden synthetische Daten. Elementar wichtig ist hier ein gutes Synthesemodell, welches unter Umständen selbst wiederum trainiert werden muss.

2.8.4 Einsatzklasse 4: Testen von Software

Als vierte Einsatzklasse fungiert das **Testen von Software**, ein in der Praxis häufig anzutreffender Vorgang. Bei der Erzeugung von Testdaten unter Berücksichtigung von Eigenschaften echter personenbezogener Daten ist zu beachten, dass die Testdaten die Echtdaten nicht derart nachbilden, dass eine Re-Identifizierung Betroffener durch Nutzung der Testdaten möglich wird. Auch die Systemmigration sowie spezifische Funktionalitätstests, bspw. von Benutzerberechtigungen, sind Anwendungsbeispiele einer Anonymisierung. Eine Anonymisierungslösung, mit der das Testen der Funktionalität von Software ermöglich wird, beinhaltet das Bereitstellen eines Testsystems, das selbst keinen Zugriff auf echte personenbezogene Daten hat.

Verantwortliche haben, neben der Bestimmung einer Rechtsgrundlage für die Anonymisierung, auch die **sonstigen rechtlichen Anforderungen der DS-GVO** einzuhalten. Hierzu zählt mit Blick auf bestehende Rechenschaftspflichten die **Dokumentation** der Anonymisierung in einem eigenen Konzept oder als Teil der Beschreibung technisch-organisatorischer Maßnahmen im Rahmen des Verzeichnisses über Verarbeitungstätigkeiten (VVT).

2.9 Sonderfall: Datenschutz-Folgenabschätzung

Aus einer Anonymisierung ergibt sich nicht per se ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen. Es müssen weitere Tatbestände hinzukommen, die in der Gesamtschau zu einem hohen Risiko führen. Liegt ein hohes Risiko vor, ist die Durchführung einer Datenschutz-Folgenabschätzung verpflichtend.

2.10 Transparenzanforderungen

Um **Transparenz** gegenüber Betroffenen zu wahren, muss bei der Datenerhebung bzw. bei der Benachrichtigung über eine geplante Anonymisierung hierüber **informiert werden**. Der Verarbeitungsvorgang der Weitergabe bereits anonymisierter Daten an Dritte unterfällt hingegen nicht mehr der DS-GVO. Über konkrete Empfänger oder Empfängerkategorien ist deshalb nicht mehr zu informieren. Auch eine Datenteilung auf Grundlage der EU-Rechtskate ist damit möglich. Sofern Daten ohne vorherige Datenschutzinformation über die beabsichtigte Anonymisierung gemäß Art. 13 DS-GVO erhoben worden sind, z.B. bei älteren Datenbeständen, erlaubt Art. 6 Abs. 4 lit. e)

DS-GVO eine Weiterverarbeitung bei Vorhandensein geeigneter Garantien. Dazu gehören nach dem Wortlaut der DS-GVO auch die Verschlüsselung oder die Pseudonymisierung. Sofern bereits eine Pseudonymisierung eine geeignete Garantie für eine zweckändernde Nutzung darstellen kann, ist die Anonymisierung eine weitaus effektivere Garantie. Ob über die nachträgliche Anonymisierung noch informiert werden muss, beurteilt sich gemäß Art. 14 Abs. 5 lit. b) DS-GVO. Diese Regelung stellt auf einen **unverhältnismäßigen Aufwand** ab. Als Beispiel hierfür werden statistische Zwecke genannt. Mit Blick auf das Persönlichkeitsrecht und den Datenschutz entstehen dem Betroffenen nach einer Anonymisierung keine Gefahren mehr. Eine nachträgliche Information über eine Anonymisierung schafft für den Betroffenen keine datenschutzrechtlichen Mehrwerte, sondern erzeugt nur einen unverhältnismäßigen Aufwand.

2.11 Prüfpflichten

Wer Daten anonymisiert hat oder anonymisierte Daten nutzt, ist verpflichtet, **kontinuierlich zu prüfen**, dass die Anonymisierung gewahrt bleibt. Dabei ist zu prüfen, ob der Personenbezug wiederhergestellt werden kann. Die Durchführung und das Prüfergebnis sollten dokumentiert werden. Auch hierbei sind Erwägungen des Angreifermodells mit einzubeziehen.

Die vorstehenden Ausführungen zeigen, dass Verantwortliche und Auftragsverarbeiter anschauliche Beschreibungen benötigen, um eine Anonymisierung personenbezogener Daten praktisch durchzuführen. Der Praxisleitfaden beschreibt daher ein Vorgehensmodell für die Anonymisierung, das gleichzeitig Ausgangspunkt für gesonderte Grundsatzregeln für die Anonymisierung darstellt. Diese wurden im Zuge dieses Praxisleitfadens ebenfalls veröffentlicht.

2.12 Vorgehensmodell für die Anonymisierung

Der Ablauf eines Anonymisierungsverfahrens lässt sich im nachfolgenden Vorgehensmodell veranschaulichen.

Lfd. Nummer	Maßnahme	Kapitel im Leitfaden
1	**Ermittlung der Rechtsgrundlage für die Anonymisierung (z.B. in der DS-GVO, im BDSG oder in einem bereichsspezifischen Gesetz)**	4.2
2	**Sicherstellen, dass die Informationspflichten gemäß Art. 13 und 14 DS-GVO umgesetzt werden**	9.3
3	**Auswahl und Festlegung des geeigneten Anonymisierungsverfahrens**	
3.1	Art und Risikoklasse der zu anonymisierenden personenbezogenen Daten	6.1 und 8
3.2	Beabsichtigte Verarbeitungszwecke	6.1 und 8

3.3	Kontext der Anonymisierung	6.1 und 8
3.4	Erwartete Anzahl der Datensätze	6.1 und 8
3.5	Ermittlung der statistischen Eigenschaften in den Datensätzen, die benötigt werden und welche Merkmale für diese Eigenschaften relevant sind	6.3 und 8
3.6	Festlegung des geeignetes Anonymisierungsverfahrens und dessen Zeitpunkt	9.1, 9.2 und 9.4
4	**Durchführung der Anonymisierung**	
4.1	Entfernung aller direkten Identifikationsmerkmale (z.B. Name, Anschrift, Kontaktdaten, Kreditkartennummer)	3.2 und 6.1
4.2	Entfernen aller nicht benötigten indirekten Identifikationsmerkmale (z.B. Geschlecht, körperliche Erscheinungsmerkmale, Alter, Postleitzahl)	3.2 und 6.1
4.3	Durchführung eines oder mehrerer Verfahren der • Randomisierung • Generalisierung oder • Durchführung eines Verfahrens mit synthetischen Daten.	6.36.3.1
5	**Analyse, ob und welche Risiken zum Wiederherstellen des Personenbezugs bestehen**	6.1 und 6.2
6	**Sofern Risiken bestehen, Anwendung weiterer Verfahren zur Anonymisierung**	6.3.1
7	**Schritte 4.1 bis 4.3 durchlaufen, bis keine Risiken mehr erkennbar sind**	6.3.1
8	**Prüfen, ob die benötigten statistischen Eigenschaften erhalten geblieben sind**	6.3
9	**Prüfung und Ergebnis dokumentieren**	9.1
10	**Anonymisierte Daten nutzen oder weitergeben**	
11	**Regelmäßig die Prüfung gemäß Schritt 5 wiederholen, ggf. Schritte 6 und 7 anwenden und gemäß Schritt 9 dokumentieren**	

3. Begriffe und Abgrenzungen

3.1 Was versteht man unter „Daten“

Der Begriff „Daten“ bezeichnet u.a. „elektronisch gespeicherte Zeichen, Angaben, Informationen“.[3] Es kommt nicht auf den Inhalt der gespeicherten Zeichen, Angaben und Informationen an. Die Beschränkung auf „gespeicherte“ Informationen in der Definition des Dudens ist unglücklich gewählt. In der praktischen Nutzung des Begriffs werden auch **Zeichen, Angaben und Informationen** als Daten bezeichnet, wenn diese zwar elektronisch verarbeitet werden, jedoch auf eine – nicht nur temporäre – Speicherung verzichtet wird.[4]

Abbildung 1: Arten von Daten

Daten liegen in **strukturierter** oder **unstrukturierter Form** vor (Abbildung 1). Strukturierte Daten haben regelmäßig die Form „Schlüssel/Key = Wert/Value“ (z.B. „Vorname = Anne“). Der Schlüssel gibt an, welche Bedeutung das Datum haben soll. In dem Beispiel wäre es „Vorname“. Der konkrete Inhalt – hier „Anne“ steht im Werteteil. Strukturierte Daten können als Schlüssel-Wert-Paar verstanden werden.

Als unstrukturierte Daten werden alle Daten bezeichnet, die nicht einer Schlüssel-Wert-Paar-Darstellung entsprechen. Dazu zählen bspw. auch Bemerkungsfelder in einem Customer Relationship Management System (CRM). In einem Bemerkungsfeld kann typischerweise beliebiger Text eingetragen werden (z.B. „Kunde ist nett“ und „bitte bei

[3] Duden Online (2022): Daten. URL: https://www.duden.de/rechtschreibung/Daten (letzter Zugriff am 28.11.2022).

[4] S. auch Art. 2 Nr. 1 Data Governance Act (DGA).

Kundenanlage auf Vollständigkeit achten“). Welche Information enthalten ist, lässt sich nicht aus der Bezeichnung „Bemerkung“ vorhersagen. Weitere Beispiele für unstrukturierte Daten sind E-Mail, Briefe, Bilder oder Tonaufnahmen.

In der Praxis kommen regelmäßig **Mischformen** aus strukturierten und unstrukturierten Daten vor. Eine E-Mail z.B. enthält als strukturierte Daten die An-Zeile, den Betreff und den Absender. Der sog. „E-Mail-Body“, d.h. der die eigentliche Nachricht enthaltende Text, ist meistens unstrukturiert.

Insbesondere bei strukturierten Daten ist es – mit Blick auf die Verfahren zur Anonymisierung – wichtig zu unterscheiden, welchen **Wertebereich** die Daten haben. Häufig vorkommende Wertebereiche sind:

- **Liste:** Es kommen nur Werte von einer definierten Liste möglicher Werte vor (z.B. „ja“ und „nein“ oder „Abteilungsleitung“, „Sachbearbeitung“, „Geschäftsführung“). Die Reihenfolge der Werte auf der Liste ist unerheblich. Enthält eine Liste nur zwei Werte, spricht man auch von einem Boolschen Wertebereich.
- **Nummerisch:** Es kommen ganze Zahlen oder Zahlen mit Kommas vor. Man spricht auch von quantitativen Daten (z.B. die Körpergröße in cm oder das Gehalt in Euro und Cent.)
- **Rating:** Das Rating ist eine Liste möglicher Werte. Die Werte stellen eine qualitative Information dar (z.B. „1 - gefällt mir“, „2 - unentschlossen“, „3 - gefällt mir nicht“). Qualitative Daten lassen sich nicht addieren o.ä. Aus „1 - gefällt mir“ plus „2 - unentschlossen“ wird nicht „3 - gefällt mir nicht“.

Strukturierte Daten werden regelmäßig in Tabellenform abgespeichert. Jede Zeile enthält die Angaben zu einem Datensatz, der bspw. eine Person repräsentiert. Die einzelnen Spalten enthalten Daten, wie z.B. Name, Alter, letzter Einkauf. Diese Daten in den Spalten werden auch Merkmale genannt. Eine solche Zeile wird auch als **„Datensatz“** bezeichnet.

3.2 Personenbezogene Daten

Personenbezogene Daten sind das Gegenstück zu anonymen Daten. Personenbezogene Daten sind gemäß der gesetzlichen Definition aus Art. 4 Nr. 1 DS-GVO

- alle Informationen
- über
- eine identifizierte oder identifizierbare natürliche Person.

3.2.1 Alle Informationen ...

Es gibt zum einen **objektive Informationen** zu einer Person wie Name, Anschrift oder Geburtsdatum. Zum anderen existieren **subjektive Informationen** wie Meinungen, Stellungnahmen oder Beurteilungen. Art und Form der Information (z.B. alphabetisch, numerisch, als Foto oder Tonaufnahme) sind dabei aus Sicht der DS-GVO irrelevant, solange sich hieraus ein inhaltlicher Aussagegehalt entnehmen lässt.

3.2.2 … über ...

Bei den Informationen muss es sich um solche über eine natürliche Person handeln. Das Vorliegen dieser Voraussetzung dient dazu, **Sachinformationen** von der DS-GVO auszuschließen. Sachdaten liegen dann vor, wenn sich eine Information nicht auf eine Person, sondern auf eine Sache bezieht („Das Kulturzentrum rote Flora befindet sich am Schulterblatt 71“).

Hinweis: Das Sachdatum darf sich weder auf eine natürliche Person beziehen, noch darf das Datum für Rückschlüsse auf eine natürliche Person verwendet werden.

Der Wert einer Immobilie ist ebenfalls ein Sachdatum, wenn es bspw. dazu verwendet wird, die Entwicklung von Immobilienpreisen in einer bestimmten Region zu beschreiben. Wird jedoch dieser Immobilienwert dazu verwendet, um den Steuersatz einer natürlichen Person zu berechnen, bezieht sich das Sachdatum auf eine natürliche Person. D.h. ein anfängliches Sachdatum kann sich zu einem späteren Zeitpunkt oder aber in einem anderen Kontext als Information über eine natürliche Person entpuppen.

3.2.3 … eine identifizierte oder identifizierbare natürliche Person

Die Information muss sich auf eine **natürliche identifizierte oder identifizierbare natürliche Person** beziehen. Eine Person ist identifiziert, wenn sie unmittelbar aus den vorhandenen Informationen selbst ermittelt werden und von anderen Personen abgegrenzt werden kann (sog. **Herausgreifen/Aussondern („singling out“)**). Das Herausgreifen/Aussondern der Person muss nicht aufgrund einer einzelnen Information erfolgen. Es können auch vorhandene Informationen kombiniert werden, um eine Person zu identifizieren.

Beispiel: Eine Kundenliste enthält Vor-, Nachname sowie Wohnort von natürlichen Personen. Bei zwei Einträgen finden sich identische Vor- und Nachnamen. Durch die Kombination von Wohnort und Vor -und Nachname lässt sich die betroffene Person direkt identifizieren.

Hinsichtlich der Identifizierbarkeit einer Person führt ErwG 26 S. 3 DS-GVO aus, dass *„**alle Mittel** zu berücksichtigen [sind], die von dem **Verantwortlichen** oder einer **anderen Person** nach **allgemeinem Ermessen wahrscheinlich** genutzt werden, um eine natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern“*.

Eine Person ist **identifizierbar**, wenn die vorhandene Information für sich betrachtet eine eindeutige Identifikation nicht ermöglicht, mittels weiterer Verarbeitungsschritte und zusätzlicher Informationen bzw. deren Verknüpfung eine betroffene Person identifiziert werden kann. Dies kann direkt über den Namen oder indirekt über eine Telefonnummer oder ein Kennzeichen erfolgen. Möglich ist auch eine Identifizierung durch Eingrenzung einer Gruppe, zu der die Person gehört.

Beispiel: Eine Liste mit Angaben zu Alter und Bruttogehalt ist im Intranet eines Unternehmens mit 35 Beschäftigten abrufbar. Bei der Beurteilung des Personenbezugs ist zu berücksichtigen, dass die Beschäftigten häufig das ungefähre Lebensalter eines Kollegen oder einer Kollegin sowie deren Position im Unternehmen kennen. D.h. mit dem Wissen um das Alter sowie der Position kann das jeweilige Gehalt erschlossen werden.

Die Zuordnung zu einer identifizierten oder identifizierbaren natürlichen Person muss nicht zutreffend sein, um von einem Personenbezug auszugehen. Auch eine **falsche Zuordnung** kann eine Information zu einer natürlichen Person darstellen. Weiterhin ist es ausreichend, wenn ein Personenbezug mit einer gewissen **Wahrscheinlichkeit** (s. hierzu Ziff. 6.1) oder nur für einen Teil der Datensätze wiederhergestellt werden kann. Sobald in einem Fall die Herstellung eines Personenbezugs gelingt, sind alle Datensätze als personenbezogen zu betrachten.

Hinweis: Ob es sich bei einer Information um eine solche über eine natürliche Person handelt, ist im Einzelfall im Hinblick auf jede einer Verarbeitung unterliegende Information unter Berücksichtigung der Sachlage und Begleitumstände zu bewerten. Für die Einstufung einer Information als personenbezogenes Datum ist es nicht erforderlich, dass die Information für sich genommen die Identifikation der betroffenen Person ermöglicht. Auch ist nicht entscheidend, ob der Name einer Person bekannt ist.

3.3 Pseudonymisierung

3.3.1 Definition und rechtliche Einordnung

Art. 4 Nr. 5 DS-GVO definiert die Pseudonymisierung[5] als

> *„die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen **nicht mehr einer spezifischen betroffenen Person zugeordnet werden können**, sofern diese **zusätzlichen Informationen gesondert aufbewahrt werden** und **technischen und organisatorischen Maßnahmen unterliegen**, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden“.*

[5] Zur Pseudonymisierung Schwartmann/Weiß (Hrsg.), Whitepaper zur Pseudonymisierung der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2017 sowie Schwartmann/Weiß, Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen - ein Arbeitspapier der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2018. Vgl. eingehend auch Schwartmann/Mühlenbeck in Schwartmann/Jaspers/Thüsing/Kugelmann, Heidelberger Kommentar DS-GVO/BDSG, Art. 4 DS-GVO Rn. 79 ff.

In der gesetzlichen Definition finden sich sowohl Hinweise zur rechtlichen Einordnung von pseudonymisierten Daten als auch zu den Anforderungen an den Vorgang der Pseudonymisierung. Klarstellend ist zu erwähnen, dass die Pseudonymisierung kein Zustand ist, sondern vielmehr ein Vorgang, der die **Umwandlung von personenbezogenen Klartextdaten in Pseudonyme** bedingt.

Werden pseudonymisierte Daten verarbeitet, handelt es sich - im Gegensatz zu anonymisierten Daten – weiterhin um **personenbezogene Daten**, die durch die DS-GVO geschützt werden. Damit erfordert die Verarbeitung pseudonymisierter Daten eine gesetzliche Rechtsgrundlage, neben den weiteren Anforderungen an die Verarbeitung personenbezogener Daten.

3.3.2 Anforderungen

3.3.2.1 Ohne Hinzuziehung zusätzlicher Informationen keine Zuordnung der Daten zu einer spezifischen Person

Pseudonymisierte Daten dürfen ohne Hinzuziehung zusätzlicher Informationen keiner Person zugeordnet werden können. D.h. bei den pseudonymisierten Daten darf es sich nicht um Informationen über eine identifizierte oder identifizierbare natürliche Person handeln. Insofern gelten die allgemeinen Abgrenzungskriterien zwischen personenbezogenen und nicht personenbezogenen und damit anonymen Daten.

Hinweis: Für eine Pseudonymisierung im Sinne der DS-GVO ist es nicht ausreichend, wenn bspw. ein reduzierter Datensatz an einen Empfänger weitergegeben wird, die Informationen aus diesem Datensatz jedoch einer natürlichen Person zugeordnet werden können.

3.3.3 Gesonderte Aufbewahrung der zusätzlichen Informationen

Pseudonymisierte Daten und vorhandene zusätzliche Informationen, die eine Re-Identifizierung eines Betroffenen ermöglichen, müssen **getrennt verarbeitet werden**. Werden personenbezogene Klartextdaten bspw. mittels kryptographischer Verfahren pseudonymisiert, hat die verantwortliche Stelle dafür zu sorgen, dass der kryptographische Schlüssel zur zum Wiederherstellen des Personenbezugs gesondert aufbewahrt wird. Eine solche Trennung kann auf **logischer Ebene** (z.B. durch Berechtigungskonzepte) aber auch auf **physikalischer Ebene** (z.B. mittels dezidierter Datenverarbeitungsanlagen) oder **organisatorischer Ebene** (z.B. über einen Datentreuhänder) erfolgen. Eine Aufteilung auf mehrere Verantwortliche wird vom Gesetzeswortlaut nicht gefordert.

Einwegfunktionen wie bspw. Hashfunktionen (s. hierzu Kapitel. 6.5.1) lassen zwar keine “Rückrechnung”, d.h. die Wiederherstellung der Klartextdaten zu. Ein Personenbezug lässt sich jedoch relativ einfach wiederherstellen, in dem die Einwegfunktion auf (vermutete) Klartextdaten angewendet wird. Ist das Ergebnis das gleiche wie beim

vorliegenden pseudonymen Datensatz, wurde die Pseudonymisierung erfolgreich aufgehoben. Deshalb sollten nur Einwegfunktionen verwendet werden, deren wiederholte Anwendung auf die gleichen Daten, zu unterschiedlichen Ergebnissen führt (bspw. Hashfunktionen mit **„Salt“** bzw. **„Pepper”**). Die ggf. vorzunehmenden Einstellungen sind analog von Zuordnungslisten getrennt aufzubewahren.

3.3.3.1 Gewährleistung technischer und organisatorischer Maßnahmen zur Nichtzuordnung

Die gesonderte Aufbewahrung von pseudonymisierten Daten und den zusätzlichen Informationen ist mit **technischen und organisatorischen Maßnahmen** zu begleiten (z.B. über ein **Berechtigungskonzept** mit unterschiedlichen technischen Rollen für einen Zugriff auf die pseudonymisierten Daten bzw. die zusätzlichen Informationen).

3.3.4 Anwendungsfälle

Typische Anwendungsfälle der Pseudonymisierung finden sich in folgenden Bereichen:

- **Forschung**
  Beispiel klinische Studie: In einer klinischen Studie werden Blutwerte von Dialysepatienten untersucht. Die Identitätsdaten werden vorab pseudonymisiert. Werden Grenzwerte in einem Blutbild überschritten, kann die betroffene Person mithilfe der Pseudonymisierungsstelle kontaktiert und um eine fachärztliche Untersuchung gebeten werden.
- **Analytics**
  Beispiel Nutzerverhalten: Ein Streaminganbieter wertet das Nutzerverhalten von seinen Kunden aus. Hierzu pseudonymisiert er die eindeutigen Gerätekennungen, um ein individuelles Nutzerverhalten zu ermitteln.
- **Werbung**
  Beispiel Datenabgleich: Ein Unternehmen pseudonymisiert seine Kundendaten, um die daraus erzeugten Pseudonyme mit denen eines Sozialen Netzwerkes abzugleichen. Hierzu wenden beide Stellen dasselbe Pseudonymisierungsverfahren an. Das Soziale Netzwerk kann nun Werbung an die Bestandskunden des Unternehmens ausspielen, ohne dass Klartextdaten aus der Kundendatenbank übermittelt werden.

3.4 Anonymisierung

Die DS-GVO enthält **keine gesetzliche Definition** für die Anonymisierung.[6] Sie ergibt sich aber im Umkehrschluss aus der Definition der „personenbezogenen Daten“ in Art. 4 Nr. 1 DS-GVO sowie aus ErwG 26 S. 3-5:

[6] Vgl. aber die europarechtlich problematische Definition in § 4 LDG NRW. Dazu Schwartmann/Mühlenbeck, in Schwartmann/Pabst Landesdatenschutzgesetz Nordrhein-Westfalen, § 4 Rn. 20 ff.

> *„[3]Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten **alle Mittel** berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach **allgemeinem Ermessen wahrscheinlich genutzt werden**, um die **natürliche Person direkt oder indirekt zu identifizieren**, wie beispielsweise das Aussondern. [4]Bei der Feststellung, ob Mittel nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind. [5]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.“*

Aus Sicht der DS-GVO ist die Anonymisierung ein **technisches Verfahren**, das auf personenbezogene Daten angewendet wird, damit natürliche Personen nicht oder nicht mehr identifiziert werden können. Die gesetzliche Definition des personenbezogenen Datums lässt offen, ob eine Identifizierbarkeit für jedermann ausgeschlossen sein muss oder ob es bspw. auf den jeweils Verantwortlichen ankommt. Auch wird nicht klar, ob der Zustand der Anonymisierung für alle Zeit zu bestehen hat. Grundsätzlich sind verschiedene Formen einer Anonymisierung denkbar.

3.4.1.1 Absolute Anonymisierung

Ist der Personenbezug praktisch für jedermann unmöglich, spricht man von **absoluter Anonymisierung**. Bei der absoluten Anonymisierung sind weder der Verantwortliche selbst noch Dritte in der Lage, die betroffene Person zu re-identifizieren. Sie ist für jedermann technisch, praktisch und faktisch, d.h. weder mit größtmöglichem Aufwand noch unter Einsatz jeglicher technischer Mittel möglich. Die absolute Anonymisierung ist **die stärkste Form** der Anonymisierung. Einen solchen Zustand zu erreichen, stellt eine große Herausforderung dar. Immerhin sind alle vorhandenen Mittel zu berücksichtigen. Dies schließt verfügbare Datenquellen im Zuge der fortschreitenden Digitalisierung ebenso mit ein, wie steigende Rechenleistungen.

Die Anzahl der Einwohner Deutschlands ist ein Beispiel für absolute Anonymisierung. Aus der Anzahl der Einwohner lässt sich nicht ableiten, ob eine bestimmte Person dazugehört oder nicht.

3.4.1.2 Faktische Anonymisierung

Die faktische bzw. relative Anonymisierung zeichnet sich dadurch aus, dass die Re-Identifizierbarkeit der betroffenen Person **nicht gänzlich ausgeschlossen ist**. Allerdings scheidet eine Re-Identifizierung der betroffenen Person aufgrund der **Unverhältnismäßigkeit ihres Aufwandes** unter Berücksichtigung der in der DS-GVO genannten sowie weiterer Kriterien aus (s. hierzu Kapitel. 6.1). In diesem Fall sind die Daten für den Verantwortlichen oder Dritten faktisch anonym.

3.4.1.3 Rechtsprechung zur Anonymisierung

Zur Streitfrage des Personenbezugs hat der Europäische Gerichtshof (EuGH) in seiner Rechtsprechung zu "**Breyer"**[7] Stellung bezogen. Hierbei hat er im Kern Folgendes festgestellt: Es ist nicht erforderlich, dass sich sämtliche, zur Identifizierung erforderlichen Informationen in der Hand eines einzigen Verantwortlichen befinden. Ausreichend ist vielmehr, wenn der Verantwortliche über einen Dritten die betroffene Person bestimmen lässt. Die Grenze der Identifizierbarkeit liegt in der Unmöglichkeit, der Unverhältnismäßigkeit oder im Rechtsverstoß. Damit hat der EuGH sich zwar im Kern einem **relativen Begriffsverständnis** angeschlossen, allerdings erhebliche Rückausnahmen zugelassen. Die Herausforderung liegt vor allem darin, dass letztlich im Rahmen der Rechtsprechung die mit Blick auf den Personenbezug entscheidende Frage offen bleibt, wann von einer Unmöglichkeit oder Unverhältnismäßigkeit der Re-Identifizierung auszugehen ist.[8]

3.4.1.4 Begriffsverständnis der DS-GVO

Mit Blick auf die in der DS-GVO enthaltenen Verhältnismäßigkeits- und Wahrscheinlichkeitserwägungen (s. ErwG 26 S. 3 u. 4) liegt es nahe, auch eine **faktische Anonymisierung** als im Einklang mit den gesetzlichen Anforderungen zu sehen. D.h. der Einsatz der Mittel und dessen Wahrscheinlichkeit wird dabei wesentlich aus Sicht des jeweils Verantwortlichen beurteilt. Ihm ist ein vorhandenes **Zusatzwissen** irgendeines Dritten nicht per se zuzurechnen. Eine Zurechnung ist jedoch geboten, wenn es sich um ein Zusatzwissen handelt, an das sich der Verantwortliche **„vernünftigerweise wenden könnte"**. Dies setzt die Kenntnis des Verantwortlichen über den Dritten und die dort vorhandenen Kenntnisse und Mittel voraus.

Ob der Einsatz **rechtswidriger Mittel** in die Prüfung der Identifizierungswahrscheinlichkeit einzubeziehen ist, wird kontrovers diskutiert. Ein vollständiges außer Acht lassen erscheint nicht sachgerecht, immerhin wird sich ein Angreifer (zum Angreifermodell s. Kapitel 6.2) mit dem Ziel einer Re-Identifizierung nicht von einem rechtswidrigen Vorgehen abschrecken lassen. Es wird in der Gesamtbewertung des Verantwortlichen darum gehen, wie wahrscheinlich und wie einfach der Einsatz solcher rechtswidriger Mittel ist.

Sollen die anonymisierten Daten an Empfänger außerhalb des Verantwortlichen **weitergegeben oder diesen Zugang eingeräumt werden**, so ist zusätzlich zu prüfen, ob die Daten auch aus der Perspektive dieser Empfänger anonym sind. Hierbei wird auf das Wissen und die Mittel der Empfänger abzustellen sein. Zu den Empfängern sind auch alle Personen oder Stellen zu zählen, die rechtlich zulässig sich die anonymisierten Daten verschaffen können bspw. über ein Auskunfts- oder Einsichtsrecht.

[7] Europäischer Gerichtshof, Urteil v. 9. Oktober 2016, C-582/14.
[8] Ausführlich hierzu Schwartmann/Mühlenbeck, RDV 2022, 264.

3.5 Anonymisierung vs. Pseudonymisierung

Die nachfolgende Darstellung verdeutlicht nochmals den Unterschied zwischen Anonymisierung und Pseudonymisierung.

Pseudonymisierung	Anonymisierung
Betroffene Person kann durch Hinzuziehung zusätzlicher Informationen wieder identifiziert werden.	Betroffene Person kann nicht oder nur mit unverhältnismäßigem Aufwand wieder identifiziert werden.
Verarbeitung ist umkehrbar.	Verarbeitung ist unumkehrbar.
Die zusätzlichen Informationen müssen gesondert aufbewahrt werden.	
Anwendungsbereich der DS-GVO ist für gesamte Verarbeitungstätigkeit eröffnet.	Anwendungsbereich der DS-GVO wird nach erfolgreicher Anonymisierung verlassen.

Der **fließende Übergang** zwischen einem anonymen und einem pseudonymen Datum soll anhand des folgenden Beispiels verdeutlicht werden:

Beispiel: Stadionbetreiber T verkauft Prepaidkarten, die ein bargeldloses Bezahlen ermöglichen. Auf den Karten ist eine zufällig generierte Kartennummer gespeichert.

Werden die Karten im Stadion ohne Angabe einer Identität erworben und aufgeladen (z.B. über die Bezahlung mit Bargeld), ermöglichen sie ein anonymes Bezahlen. Es handelt sich daher bei der Kartennummer um ein anonymes Datum.

Wenn Kunden die Karten jedoch zusätzlich online aufladen können, was eine Registrierung mit personenbezogenen Daten und eine Verknüpfung mit der Prepaidkarte erforderlich macht, kann die Identität der Kunden mittels der zusätzlichen Informationen ermittelt werden. Für den Stadionbetreiber ist die Kartennummer nunmehr nicht als anonymes Datum, sondern vielmehr als Pseudonym einzustufen.

3.6 Künstliche Intelligenz

Ein Computerprogramm, auch **Algorithmus** genannt, stellt - vereinfacht ausgedrückt - einen Satz von Regeln im Stile von „wenn x zutrifft, dann tue y“ und „mache das ganze z-Mal“ dar. Insofern unterscheidet sich ein Computerprogramm nicht von einem Kochrezept oder einer Klickanleitung. Die im Computerprogramm eingebauten Regeln geben das Verständnis der Entwickler von dem im Computerprogramm bearbeiteten Sachverhalt wieder. Dieses Verständnis kann - muss aber nicht - wissenschaftlich fundiert, richtig oder gar ethisch erwünscht sein.

Auch die „**Künstliche Intelligenz“ (KI)** ist nichts anderes als ein solches Computerprogramm. Sie hat nichts – absolut gar nichts – mit „Intelligenz“ im menschlichen Sinn zu tun. Ehe der Begriff „Künstliche Intelligenz“ zu einem Modewort wurde, das überlegene und moderne Technik suggerieren soll, bezeichnete er lediglich ein Teilgebiet der Infor-

matik. Dieses Teilgebiet beschäftigt sich mit ganz unterschiedlichen **Klassen von Algorithmen**. Deshalb gibt es auch nicht „die" KI, sondern eine Vielzahl von unterschiedlichen Arten von Algorithmen und Vorgehensweisen, die jeweils anders funktionieren und unterschiedliche Einsatzgebiete haben.

Es handelt sich im Kern nicht um eine neue oder moderne Technik. Das Teilgebiet „Künstliche Intelligenz" entstand 1956[9]. In den vergangenen 66 Jahren wurden die Methoden und Algorithmen stetig weiterentwickelt und leistungsfähiger gemacht.

Zu der heute in der öffentlichen Wahrnehmung domminierenden Algorithmen-Klasse zählt die **Mustererkennung mittels maschinellen Lernens** (kurz „Mustererkennung" im Folgenden). Teilweise wird im Marketing, in der Gesetzgebung, (juristischen) Literatur sowie in der Öffentlichkeit KI mit Mustererkennung gleichgesetzt.

Algorithmen zur Mustererkennung zeichnen sich dadurch aus, dass in der Phase der Programmierung des Algorithmus weder die zu erkennenden Muster noch die Merkmale, anhand derer sich die Muster erkennen lassen, bekannt sein müssen. Die Algorithmen zur Mustererkennung enthalten unspezifische Regeln, die erst in einer **„Trainingsphase"** angepasst werden, so dass mit einer gewissen Wahrscheinlichkeit das von den Entwicklern als „richtig" betrachtete Muster erkannt wird.

Das Training verläuft - vereinfacht dargestellt - wie folgt:

1. Dem Algorithmus werden verschiedene Daten, zu denen auch Bilder zählen, als Eingabe präsentiert.
2. Der Algorithmus „würfelt" die Lösung.
3. Bei einer richtigen Lösung erhält der Algorithmus eine „Belohnung" und bei einer falschen Lösung wird er bestraft.
4. Passe an, wie Du würfelst
5. Fange mit dem nächsten Datensatz von vorne an.

Nach teilweise Millionen Trainingsdurchläufen erkennt der Algorithmus die Trainingsmuster mit einer **zufriedenstellenden Wahrscheinlichkeit**. Welche Muster er erkennen kann, wird durch die Trainingsdaten und seine Programmierung bestimmt, d.h. er erkennt immer nur bestimmte, vom Hersteller des Produktes, in dem der Algorithmus arbeitet, festgelegte Muster. Das Training ist an dieser Stelle in der Regel beendet, d.h. im produktiven Einsatz findet kein „Lernen" mehr statt. Ausnahmen von dieser Regel bestehen je nach Einsatzzweck.

Erkennen bedeutet nicht, dass der Algorithmus das Muster so wie ein Mensch semantisch versteht. Aus Sicht des Algorithmus ersetzt er das Muster durch ein Wort.

Der grobe Ablauf lässt sich am Beispiel einer Bilderkennung wie folgt vorstellen:

1. Identifiziere (vermutlich) zusammengehörige Punkte im Bild (z.B. die gleiche Farbe)
2. Betrachte zusammengehörige Punkte als „Objekt x"

[9] McCarthy et al. (1955): A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Förderantrag, August 1955, S. 1. URL: https://web.archive.org/web/20080930164306/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, letzter Zugriff am 28.11.2022.

3. Durchsuche Liste der bekannten Objekte auf Übereinstimmung mit Objekt x
4. Gebe das Wort y aus, das dem Objekt x in der Objektliste zugeordnet ist

Aus Sicht eines Algorithmus zur Mustererkennung sieht ein Bild wie eine Wolke von farbigen Punkten aus. Er „sieht“ keine Linien oder geometrische Figuren wie ein Mensch. Erst recht versteht er den Sinn des Bildes nicht.

Das Schöne an einem Algorithmus ist, dass er immer eine Antwort liefert. „Weiß nicht“ gehört anders als beim Menschen nicht zu seinem Sprachschatz. Deshalb entsteht beim menschlichen Betrachter leicht der Eindruck, dass die Antwort eines Algorithmus zur Mustererkennung „richtig“ sei. Dieser Eindruck täuscht, da das „Erkennen“ nach menschlichen Verständnis ein „Raten“ ist. Die Antworten sind nicht immer richtig. Anders ausgedrückt: Die erkannten Muster sind im besten Fall **häufiger richtig als falsch**. Wer solche Algorithmen einsetzt, muss also mit falschen Ergebnissen rechnen und sich überlegen, wie er damit umgeht. Je nach Verwendung führen Erkennungsfehler zu Toten.[10]

Man misst die Güte des trainierten Mustererkennungsalgorithmus anhand der Maße

- **Anteil False Positive:** Ein Bild zeigt ein Pferd, der Algorithmus erkennt jedoch eine Ente
- **Anteil False Negative:** Ein Bild zeigt einen Schwarm, der Algorithmus erkennt jedoch ein Pferd

Auch bei richtigen Antworten eines Algorithmus zur Mustererkennung bedeutet das nicht, dass der Algorithmus auf die gleichen Merkmale achtet wie ein Mensch.

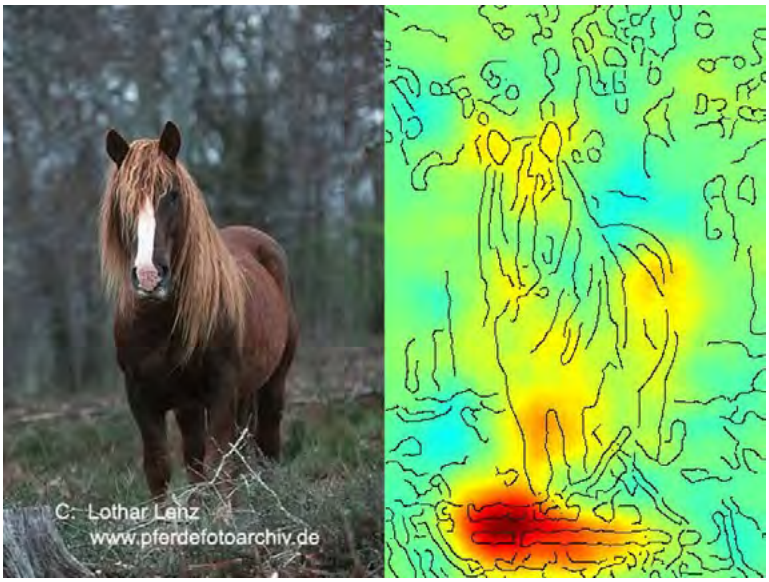

Abbildung 2: Das rechte Bild zeigt in rot die analysierten Bildbereiche des linken Bildes[11]

[10] Vgl. bspw. Der Standard (2022): "Full Self-Driving": Erneut tödlicher Tesla-Unfall mit aktiviertem Autopiloten. 03.08.2022, URL: https://www.derstandard.de/story/2000137996067/full-self-driving-erneut-toedlicher-tesla-unfall-mit-aktiviertem-autopilot (letzter Zugriff am 28.11.2022).

[11] Quelle: Heise Online (2020): Wie sich KI-Entscheidungen überprüfen lassen. URL: https://heise.de/-4665982 (letzter Zugriff am 28.11.2022).

Das rechte Bild in Abbildung 2 visualisiert die Bildabschnitte des linken Bildes anhand derer ein Pferd erkannt wird. Es wird der Copyright-Hinweis verwendet, da Trainingsbilder mit Pferden sich durch den Copyright Hinweis von anderen Bildern unterschieden.[12] Würde dem Algorithmus nun ein Bild eines Hauses mit dem gleichen Copyright Hinweis gezeigt, würde der Algorithmus das Haus als Pferd erkennen.

Weil die Regeln, Muster zu erkennen nicht durch den Softwareentwickler explizit in den **Quellcode** des Algorithmus geschrieben werden, sondern sich erst in der Trainingsphase „herausbilden", lässt sich nicht ohne Hilfsmittel nachvollziehen, wie ein Ergebnis des Algorithmus zustande kommt. Bei Algorithmen, die ohne maschinelles Lernen entwickelt werden, stehen die Regeln explizit und im Prinzip für einen Menschen einfach wie ein Buch lesbar im Quellcode. Die Funktionsweise lässt sich – genügend Zeit und Wissen vorausgesetzt – von jedem Menschen durch Lesen des Quellcodes nachvollziehen.

Bei Algorithmen, die mittels maschinellen Lernens entwickelt werden, besteht eine solche Möglichkeit nicht. Es bedarf zusätzlicher technischer Hilfsmittel, die Funktionsweise transparent zu machen. Aus diesem Grund werden Fehler häufig, wenn sie überhaupt erkannt werden, erst im produktiven Einsatz gefunden. Teilweise mit fatalen Folgen für die beteiligen Menschen.

4. Rechtliches Umfeld

4.1 Anonymisierung im Lichte der europäischen Datenstrategie

Datenverarbeitung wird in der EU künftig in einem größeren Kontext eingebettet sein. Die zentralen Rechtsakte sind der schon verabschiedete **Data Governance Act (DGA)** und der entstehende **Data Act (DA)**. Nach ersterem bekommen öffentliche Stellen die Möglichkeit, Daten zur Weiterverwendung bereitzustellen. Letzterer adressiert die Wirtschaft. Daten, die sich aktuell in den Händen von Herstellern von **IoT-Geräten** befinden, sollen auch für Nutzer und andere Unternehmen wirtschaftlich verwertbar gemacht werden. Hierzu werden Nutzer befähigt, ihre Daten aus solchen Geräten teilen zu können. Auf diese Weise will der Gesetzgeber Anreize für innovative Geschäftsideen an der richtigen Stelle schaffen. Ergänzend erlegt der **Digital Markets Act (DMA)** den „Gatekeepern", welche die Digitalwirtschaft auch innerhalb der EU dominieren, Pflichten auf, um fairen Wettbewerb im Binnenmarkt herzustellen. Nutzer sollen mehr Datensouveränität erhalten. Der **Digital Services Act (DSA)** wiederum beansprucht nicht weniger, als die Demokratie zu sichern. Er verpflichtet die großen Online-Plattformen insbesondere dazu, Hass, Fakenews und Kriminalität im Netz zu bekämpfen. Die Konzerne müssen Verfahren etablieren, die Risiken ihres Geschäftsmodells mindern. Besondere Bedeutung misst die EU auch der **Verordnung für Künstliche Intelligenz (KI)** bei – sie soll 2023 verabschiedet werden und die EU zum weltweiten Trendsetter einer fairen Nutzung dieser Schlüsseltechnologie machen. Die Aktivitäten der EU erstrecken sich

[12] Heise Online (2020): Wie sich KI-Entscheidungen überprüfen lassen. URL: https://heise.de/-4665982 (letzter Zugriff am 28.11.2022); Weiterführende Literatur: Christopher J. Anders, Talmaj Marinc et al. (2019): Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed arXiv:1912.11425, 2019.

schließlich auch auf die Sicherheit des Internets. Der Entwurf zur Überarbeitung der **Richtlinie zur Netz- und Informationssicherheit (NIS)** sieht eine deutliche Ausweitung des ursprünglichen Anwendungsbereichs vor und erfasst nun auch kleinere Unternehmen. Hier werden ebenso neue Pflichten festgelegt wie in der **Radio Equipment Directive (RED)**. Diese adressiert Gefahren vernetzter Endgeräte (IoT) und muss bis zum Jahr 2024 umgesetzt werden. Geplant ist zudem ein **Cyber Resilience Act (CRA)**, der die technische Stabilität für im Netz eingesetzter Produkte gewährleisten soll.

Die in den neuen Rechtsakten vorgesehenen Mechanismen für das Datenteilen warten zumeist mit der Möglichkeit zur Implementierung von Schutzmaßnahmen zugunsten Betroffener auf. Hierzu gehören auch die **Pseudonymisierung** und **Anonymisierung** personenbezogener Daten. Die Anonymisierung von Daten wird also mit den neuen Datenakten signifikant an Bedeutung gewinnen. Die DS-GVO bleibt der **Standard**, wenn es um die Verarbeitung personenbezogener Daten geht, der zusätzlich zu spezifischen Vorgaben der neuen Rechtsakte einzuhalten ist. Gesetze in der EU, die den Schutz personenbeziehbarer Daten betreffen, lassen die DS-GVO nämlich unberührt. Damit gelten die Aussagen der DS-GVO weiterhin für die wichtige Unterscheidung zwischen einer Pseudonymisierung und Anonymisierung personenbezogener Daten.

4.2 Anonymisierung als Datenverarbeitung im Sinne der DS-GVO

Anders als für pseudonymisierte Daten gilt die DS-GVO nicht für anonyme Daten. Die Anonymisierung zeichnet sich jedoch dadurch aus, dass mithilfe einer Technik aus personenbezogenen Daten anonyme Daten werden. Die zunächst erhobenen personenbezogenen Daten müssen im Einklang mit der DS-GVO, so auch unter Beachtung der Anforderungen einer **Rechtsgrundlage**, verarbeitet werden. Ob es für die Anonymisierung selbst auch einer Rechtsgrundlage bedarf, ist anhand des Anwendungsbereichs der DS-GVO zu beurteilen. Sie ist anzuwenden, wenn personenbezogene Daten ganz oder teilweise automatisiert verarbeitet werden oder eine nichtautomatisierte Verarbeitung erfolgt und die Daten in einem Dateisystem gespeichert werden. Konsultiert man die Definition der **Verarbeitung** in Art. 4 Nr. 2 DS-GVO, liegt es nahe, das Überführen des Status „personenbezogen“ in „anonym“ regelmäßig als Verarbeitung in Form einer Veränderung personenbezogener Daten anzusehen. Damit sind für den Vorgang der Anonymisierung die rechtlichen Anforderungen der DS-GVO zu beachten, so auch das Vorliegen einer Rechtsgrundlage (zu den sonstigen rechtlichen Anforderungen vgl. Kapitel 9).

Die **Tabelle** gibt einen Überblick über üblicherweise einschlägige Rechtsgrundlagen. Welche Rechtsgrundlage eine konkrete Anonymisierung erlaubt, muss im Einzelfall beurteilt werden.

Datenart	Zweck „Anonymisierung“ bei Erhebung mitgeteilt	Zweck	Übliche Rechtsgrundlage in der DS-GVO
-/-	-/-	Löschung	Rechtliche Verpflichtung (Art. 6 Abs. 1 lit. c) i.V.m. 17 Abs. 1 lit. a))
„normal“	ja	Weitergabe an Dritte	• Einwilligung (Art. 6 Abs. 1 lit. a)) • Interessenabwägung (Art. 6 Abs. 1 lit. f) – nur bei nicht öffentlichen Stellen)
„normal“	nein	Weitergabe an Dritte	Prüfen, ob neuer Zweck kompatibel mit ursprünglichem Zweck der Verarbeitung ist (Art. 6 Abs. 4)
Besondere Kategorien (Art. 9 Abs. 1 DS-GVO)	ja	Weitergabe an Dritte	• Einwilligung (Art. 9 Abs. 1 lit. a)) • Rechtliche Verpflichtung aus Arbeits- oder Sozialrecht (Art. 9 Abs. 1 lit. b))
Besondere Kategorien (Art. 9 Abs. 1 DS-GVO)	nein	Weitergabe an Dritte	Nicht zulässig
Strafrechtliche Verurteilungen & Straftaten (Art. 10)	-/-	Weitergabe an Dritte	Grundsätzlich unzulässig

Tabelle 1: Überblick üblicherweise anwendbarer Rechtsgrundlagen[13]

[13] Lepperhoff, N. (2022): Anonymisierung von personenbezogenen Daten – Teil 2: Praxisbeispiele und Randbedingungen. In: Lohn und Gehalt 07/2022.

5. Funktionen der Anonymisierung

Eine Anonymisierung verhindert, dass sich Daten bestimmten Personen zuordnen lassen. Hierfür **entfernt, ersetzt, aggregiert oder verfälscht** der Vorgang des Anonymisierens personenbezogene Daten oder personenbeziehbare Daten (siehe Kapitel 6.3.1). Bei anonymisierten Daten wird damit eine Re-Identifikation verhindert. Für vollständig anonymisierte Daten gelten die Vorgaben der DS-GVO nicht mehr (s. Kapitel 3.4). Der Verantwortliche ist demnach an datenschutzrechtliche Vorgaben zur Zulässigkeit der Datenverarbeitung anonymisierter Daten nicht mehr gebunden. Damit ist eine weitere Nutzung zur Analyse oder zur Weitergabe grundsätzlich zulässig. Inwieweit sich die Anonymisierung für erwünschte Einsatzzwecke eines Verantwortlichen eignet, hängt wesentlich von der eingesetzten **Technik** und einer erwünschten **Nutzbarkeit von Daten** ab (zu ausgewählten Einsatzklassen s. Kapitel 8).

Hinweis: Ein Verantwortlicher ist nicht daran gehindert, die Anonymisierung als **technisch-organisatorische Maßnahme** einzusetzen, um das Risiko für Rechte und Freiheiten Betroffener zu mindern. Gerade in Situationen, in denen Unsicherheit bezüglich des Risikos einer Re-Identifizierung besteht, kann die Anonymisierung weiterhin sinnvoll sein. Der Anwendungsbereich der DS-GVO würde dann zwar nicht verlassen werden, personenbezogene Datenverarbeitungen können durch diese Maßnahme jedoch ermöglicht werden. Ebenso hat die Anonymisierung Relevanz beim Gebot der **Datenminimierung** gem. Art. 5 Abs. 1 lit. c) DS-GVO.

6. Anforderungen an die Anonymisierung

6.1 Rechtlich

Die DS-GVO enthält keine konkreten Anforderungen, wann personenbezogene Daten hinreichend anonymisiert worden sind. Die Definition des personenbezogenen Datums aus Art. 4 Nr. 1 DS-GVO sowie die Erläuterungen in ErwG 26 DS-GVO (zu den personenbezogenen Daten vgl. Kapitel 3.2) enthalten jedoch einige Anhaltspunkte und geben Hinweise, welche Anforderungen an die Anonymisierung aus rechtlicher Sicht zu stellen sind.

Mit Blick auf Art. 4 Nr. 1 DS-GVO darf der aus einer Anonymisierung entstandene Datensatz keine Informationen über eine identifizierte oder identifizierbare Person enthalten.

Ob sich die die aus einer Anonymisierung erzeugten Daten auf eine identifizierte oder identifizierbare Person beziehen, wird in ErwG 26 S. 3 u. 4 DS-GVO konkretisiert:

> *„Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten **alle Mittel** berücksichtigt werden, die von dem V**erantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden**, um die natürliche Person **direkt** oder **indirekt** zu identifizieren, wie beispielsweise das **Aussondern**. Bei der Feststellung, ob **Mittel nach allgemeinem Ermessen***

> ***wahrscheinlich*** *zur Identifizierung der natürlichen Person genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologischen Entwicklungen zu berücksichtigen sind."*

Der Stelle, die eine Anonymisierung ausführt, wird durch en Erwägungsgrund zunächst eine **Prüfpflicht** auferlegt, ob es sich nach dem Vorgang um personenbezogene Daten handelt oder nicht. Im Rahmen dieser Prüfung sind **alle Mittel** zu berücksichtigen, die **vernünftigerweise** entweder von dem Verantwortlichen oder einem Dritten eingesetzt werden könnten, um die betroffene Person zu identifizieren. Solche Mittel können bspw. für den Verantwortlichen verfügbare Informationen sein, oder solche, die er sich beschaffen kann. Insofern wird er auch die **Verknüpfung** sog. indirekter Identifikationsmerkmale berücksichtigen müssen, die zu einer Identifizierung eines Betroffenen führen kann. Auch **Kontextinformationen** oder **Rohdaten** können eine Rolle spielen. Gerade mit Blick auf die europäische Datenstrategie sind öffentlich-zugängliche Datenräume mit personenbezogenen, pseudonymisierten oder anonymisierten Daten vermehrt zu erwarten. Eine Vielzahl von Datenquellen kann die **Wahrscheinlichkeit einer Re-Identifizierung** erhöhen. Irrelevant ist, ob der Verantwortliche oder ein Empfänger der Daten die Person identifizieren möchte oder nicht. Es reicht eine **objektive Identifizierbarkeit** aus.

Es gibt Fälle, in denen die Wahrscheinlichkeit des Mitteleinsatzes nicht ohne Weiteres beantwortet werden kann. Hier benennt die DS-GVO einige **Faktoren**, die als Prüfkriterien mit Blick auf das jeweils vorhandene Mittel beim Verantwortlichen oder beim Dritten verwendet werden können:

- Kosten der Identifizierung
- Erforderlicher Zeitaufwand
- Verfügbare Technologien zum Zeitpunkt der Verarbeitung und deren Entwicklung
- Sonstige objektive Faktoren

Aus technischer Sicht wird sich ein Verantwortlicher die Frage stellen müssen, ob es zum Zeitpunkt der Verarbeitung, sprich der Anonymisierung, Technologien gibt, die eine Re-Identifizierung begünstigen können. Entsprechend hat die anonymisierende Stelle auch ein Verfahren nach dem **Stand der Technik**[14] einzusetzen. Da Technologien nicht statisch bleiben, sind auch deren **Entwicklungen** mit in die Analyse einzubeziehen. Löscht ein Verantwortlicher beispielsweise den Schlüssel zum Verschlüsseln eines Da-

[14] "Stand der Technik ist der Entwicklungsstand fortschrittlicher Verfahren, Einrichtungen und Betriebsweisen, der nach herrschender Auffassung führender Fachleute das Erreichen des gesetzlich vorgegebenen Zieles gesichert erscheinen lässt. Verfahren, Einrichtungen und Betriebsweisen oder vergleichbare Verfahren, Einrichtungen und Betriebsweisen müssen sich in der Praxis bewährt haben oder sollten – wenn dies noch nicht der Fall ist – möglichst im Betrieb mit Erfolg erprobt worden sein." (vgl. Handbuchs der Rechtsförmlichkeit v. 22.09.2008, Rn. 256).

tensatzes, sollte der Datensatz nicht für alle Zeit als anonym eingeschätzt werden. Immerhin kann der technologische Fortschritt (Stichwort: **Quantencomputing**) zu einer Umkehr der Verschlüsselung in Zukunft sorgen.[15]

Hat der Verantwortliche die **rechtliche Möglichkeit**, mittels Ausübung von Rechten eine Person zu (re-)identifizieren, ist dies in die Wahrscheinlichkeitsprüfung mit einzubeziehen.

Das **Interesse an einer Re-Identifizierung** spiegelt sich regelmäßig im Wert der Daten für einen Angreifer wider. Auch dies kann in eine Wahrscheinlichkeitsprüfung mit einbezogen werden. So ist die Umkehrung anonymisierter Kreditkartensätze für einen Angreifer von großem Interesse, was Auswirkungen auf die Stärke der Anonymisierung haben muss. Auch können Gesundheitsdaten für einen Angreifer von besonderem Interesse sein, bspw. um die Erkrankung einer bestimmten Person zu bestimmen (zum Angreifermodell s. Kapitel 6.2).

Ob anonymisierte Daten an **interne Empfänger** weitergegeben werden oder solche Daten **veröffentlicht** werden, sollte Auswirkungen auf die Risikoanalyse einer Re-Identifizierung haben. Bei einer Veröffentlichung besteht regelmäßig ein großer **Empfängerkreis**, der mittels Zusatzinformationen eine betroffene Person ggf. re-identifizieren kann. Daher sollten bei einer Veröffentlichung von anonymisierten Daten strenge Maßstabe bei besagter Risikoanalyse gelten.

Es können zu einem **späteren Zeitpunkt** auch neue Mittel in Betracht zu ziehen sein, deren Vorliegen einen Einfluss auf die Frage einer hinreichenden Anonymität haben kann. Anonymisierungsverfahren sind daher fortlaufend zu **überprüfen** und zu e**valuieren** (zu den Prüfpflichten vgl. Kapitel 9.4).

Aus Sicht der DS-GVO können, mangels gesetzlicher Präzisierung, verschiedene **Anonymisierungstechniken** eingesetzt werden. Entscheidend ist, dass nach Prüfung der

> **Hinweis:** Das Entfernen von **direkten Identifikationsmerkmalen** (z.B. der Name einer Person) genügt in vielen Fällen noch nicht, um personenbezogene Daten zu anonymisieren. Auch aus weiteren, zur Verfügung stehenden Informationen kann eine Person möglicherweise identifiziert werden (z.B. Geschlecht, Berufsgruppe und Geburtsjahr). Kann der Verantwortliche selbst oder ein Dritter ohne unverhältnismäßig großen Aufwand Betroffenene re-identifizieren oder identifizierbar machen, sind personenbezogene Daten per se nicht hinreichend anonymisiert.

oben aufgeführten Faktoren eine Re-Identifizierung von Betroffenen praktisch nicht durchführbar ist. D.h. erfordert sie einen unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft, kann grundsätzlich von einer wirksamen Anonymisierung ausgegangen werden.

[15] Vgl. EDPS und Agencia Española de Protección de Datos, 10 Misunderstandings related to Anonymisation, abrufbar unter https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf (letzter Zugriff am 28.11.2022).

Solange der zur Anonymisierung verwendete personenbezogene Datensatz beim Verantwortlichen vorhanden ist, bleiben auch die hieraus anonymisierten Daten für ihn regelmäßig personenbezogen. Damit finden die DS-GVO und alle übrigen Datenschutzgesetze auch auf den anonymisierten Datenbestand Anwendung, sollte die Anonymisierung durch den Zugang zum ursprünglichen Datensatz ohne Weiteres aufhebbar sein.

Hinweis: Ein Verantwortlicher wird aus rechtlicher Sicht die im Gesetz benannten Faktoren sowie alle weiteren Umstände in einer **Gesamtschau** dahingehend prüfen müssen, wie wahrscheinlich es ist, dass jemand den Aufwand einer Re-Identifizierung betreiben wird.

6.2 Angreifermodell

Um zu beurteilen, ob Daten anonym sind, bietet es sich an, die Perspektive **eines „Angreifers“** einzunehmen, der versucht einen Personenbezug wiederherzustellen oder Aussagen über eine konkrete Person aus den Daten abzuleiten.
Ein **„Angreifermodell”** beschreibt damit eine Methode, mit der geprüft wird, ob ein Datensatz anonym oder personenbezogen ist.[16] Aus der Perspektive eines Angreifers wird getestet, ob eine Re-Identifikation möglich ist. Erst wenn ein solcher – ernsthaft durchgeführter – Versuch scheitert, lässt sich von anonymen Daten sprechen.

Welche **Kenntnisse** und **Fähigkeiten** dem Angreifer unterstellt werden, hängt vom **Verwendungskontext** der Daten ab. Sollen die Daten veröffentlicht werden, ist ein – mit Blick auf die Vielzahl von (kriminellen und staatlichen) Akteuren - von tieferen Fachwissen und einer höheren Ausstattung auszugehen, als wenn die Daten an einen bestimmten Empfänger gegeben werden. Ein Angreifer kann verschiedene Ziele verfolgen, um Personen in einem anonymisierten Datensatz zu re-identifizieren Je wertvoller die Daten für den Angreifer sein können, desto mehr Fachwissen und Ressourcen sind zu unterstellen. Dabei sind nicht alle theoretisch denkbaren oder nicht ausschließbaren technischen Möglichkeiten oder möglicherweise vorhandenes Wissen einzubeziehen, sondern die vernünftigerweise wahrscheinlichen.[17]

Es sind grundsätzlich verschiedene **Arten eines Angriffs** denkbar:

- Ein Angriff, dem ein **vorhandenes Wissen** über eine Person zugrunde liegt und von der bekannt ist, dass sie in einem Datensatz existiert.
- Ein Angriff, der sich **öffentliche oder andere ihm zur Verfügung stehende Quellen** zunutze macht, um die Person zu re-identifizieren.

[16] Art- 29 Gruppe (2014): Stellungnahme 5/2014 zu Anonymisierungstechniken. Angenommen am 10. April 2014. WP216, ICO (2021): draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, Kapitel 2, S. 14ff.
[17] ICO (2021): draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, Kapitel 2, S. 12.

- Ein Angriff, der darauf abzielt, **so viele Personen wie möglich** aus dem Datensatz zu re-identifizieren, auch wenn dies bedeutet, dass Personen fälschlicherweise identifiziert werden.

Hinweis: Es empfiehlt sich, beim Angreifermodell nicht nur **zielgerichtete** Angriffe zu berücksichtigen, sondern auch Konstellationen, in denen eine Re-Identifizierung durch den Angreifer eigentlich **ungewollt ist bzw. zufällig** von statten gehen könnte. Ebenso kann es Angriffe geben, die darauf abzielen mit den betroffenen Personen zu „interagieren", ohne dabei zu wissen, um wen es sich konkret handelt (z.B. indem zusätzliche Informationen zu einer Person aus dem Datensatz erlangt werden sollen).

Es bestehen verschiedene **Umsetzungsmöglichkeiten**, um einen Angriff mit dem Ziel einer Re-Identifizierung durchzuführen:

- **Herausgreifen/Aussondern („singling out")**: Ausgewählte Datensätze werden aus einem Datenbestand herausgegriffen, um eine Person zu identifizieren. Die betroffene Person kann aufgrund der im Datensatz vorhandenen oder dem Angreifer zugänglichen Informationen hinreichend von anderen Personen unterschieden werden. Bspw. kann in einer Gehaltsliste der Geschäftsführer die Person mit dem höchsten Gehalt sein.

  Beispiel: Ein Angreifer leitet aufgrund unzureichender Diversität in einem Datensatz, hier bezüglich sensibler indirekter Identifikationsmerkmale (z.B. „Person X hat Krebs"), die Zuordnung einer Person zu einem solchen Merkmal ab. Ein solcher Angriff wird begünstigt, wenn es nicht genügend Variationen von sensiblen Merkmale in einer Gruppe von Personen gibt (s. auch k-Anonymität unter Ziff. 6.3.1.2.1 und l-Diversität unter Ziff.6.3.1.2.2).

- **Verknüpfung von Datensätzen ("record linkage")**: Der Angreifer versucht, einen Datensatz aus der anonymisierten Datenbank mit einer Person zu verknüpfen. Mit Hintergrundwissen über mittelbare Identifikatoren kann der Angreifer eine kleine Gruppe von Datensätzen oder möglicherweise einen einzelnen Datensatz mit einer Person verknüpfen. Die Verknüpfung kann bspw. auch mit statistischen Verfahren erfolgen. Es reicht für die Feststellung des Personenbezugs aus, dass eine Wahrscheinlichkeit besteht, dass zwei Datensätze zur gleichen Person gehören.
- **Inferenz**: Ein (neues) Merkmal einer Person wird den Werten anderer im Datenbestand vorhandener Merkmale abgeleitet. Dabei wird mit statistischen Verfahren nach Beziehungen zwischen Merkmalen gesucht. Für den Personenbezug reicht es aus, wenn die vermutete signifikant wahrscheinlich ist.

  Hinweis: In das Angreifermodell sind alle Datenbestände einzubeziehen, auf die der Verantwortliche, Empfänger der anonymisierten Daten sowie der Angreifer Zugriff hat oder haben kann.

6.3 Technisch

Anonymisierungsmethoden **wenden Datentransformationen auf die Originaldaten** an, um die gewünschten Eigenschaften des anonymen Datensatzes zu erreichen. Das heißt, **Identifikatoren** (sowohl direkte als auch indirekte) werden entfernt und/oder umgewandelt, so dass ein Datensatz mit Informationen nicht mehr einer bestimmten Person zugeordnet werden kann.

Anonymisierungsverfahren bewegen sich auf einem schmalen Grad. Einerseits soll der Personenbezug aufgehoben werden. Anderseits sollen die **statistischen Eigenschaften** der Daten nicht verändert werden. Ein Anonymisierungsverfahren verändert notwendigerweise die Daten, um den Personenbezug aufzuheben. Die entscheidende Frage ist, ob die Veränderung auch zu einer unerwünschten Veränderung der statistischen Eigenschaften führt.

Nicht jedes Anonymisierungsverfahren ist für jeden Einsatzfall geeignet. Wird ein ungeeignetes Verfahren verwendet, kann der Personenbezug bestehen bleiben oder der Personenbezug wird zusammen mit den benötigten statistischen Eigenschaften entfernt. Das Ergebnis wäre im besten Fall nutzlos.

Bei der Auswahl eines oder mehrere Anonymisierungsverfahren sollte der Anwender feststellen, welche statistischen Eigenschaften seiner Daten auf jeden Fall erhalten bleiben sollen. Das bedeutet auch, die in den Daten enthaltenen statistischen Eigenschaften zu ermitteln und fachlich zu verstehen. Ein solches Verständnis hängt vom Anwendungsgebiet und den konkreten Daten ab.

Häufig müssen **mehrere Anonymisierungsverfahren** hintereinander ausgeführt werden. Welche das sind, und wie die Reihenfolge aussehen könnte, hängt ebenfalls vom Einzelfall ab.

An dieser Stelle werden deshalb die Anonymisierungsverfahren abstrakt vorgestellt, um einen Eindruck zum Anwendungsbereich und zur Funktionsweise zu vermitteln. Ob für konkrete Anwendungsfälle fertige Produkte oder Open-Source-Implementationen auf dem Markt erhältlich sind, wäre im Einzelfall zu prüfen. Ob ein Produkt geeignet ist, sollte sorgfältig geprüft werden. Ist kein Produkt vorhanden, bleibt nur die eigene softwaretechnische Implementierung. Weiterhin stellt die Übersicht eine Momentaufnahme dar, da an Anonymisierungsverfahren fortlaufend geforscht wird.

Anonymisierungsverfahren lassen sich in zwei Klassen einteilen:
- Randomisierung
- Generalisierung

Im Folgenden werden bekannte Verfahren vorgestellt, die vorhandene personenbezogene Daten anonymisieren können.[18] Im Bereich des **maschinellen Lernens** sind weitere Verfahren bekannt, die zwar keine Daten anonymisieren, jedoch die Eingriffstiefe

[18] Die Darstellung basiert im Wesentlichen auf Art- 29 Gruppe (2014): Stellungnahme 5/2014 zu Anonymisierungstechniken. Angenommen am 10. April 2014 (WP 216).

in das Persönlichkeitsrecht im Trainingsprozess oder Trainingsergebnis reduzieren. Diese Verfahren werden in Kapitel 8.3 beispielhaft vorgestellt.

6.3.1 Ausgewählte Verfahren

6.3.1.1 Verfahren der Randomisierung

Im Kern werden bei der **Randomisierung** Werte zufällig verändert. Diese Veränderung führt zum Aufheben einer Verbindung zwischen verschiedenen Merkmalen. Es werden Inferenzrisiken reduziert. Die Art und Weise der Veränderungen hängt vom gewählten Verfahren ab.

6.3.1.1.1 Stochastische Überlagerung

Die **stochastische Überlagerung** verändert die Werte einzelner Merkmale in einem Datensatz. Voraussetzung ist, dass die Werte **nummerisch**, d.h. quantitativ sind. Weiterhin wird vorausgesetzt, dass die Originaldaten nach Anwendung des Verfahrens gelöscht werden, damit die Veränderung nicht nachvollzogen werden kann.

Die Veränderung erfolgt nicht willkürlich, sondern so, dass die statistische Verteilung der ursprünglichen Werte nicht verändert wird. Kommt der ursprüngliche Wert in 5 von 100 Fällen vor, so soll der veränderte Wert ebenfalls in 5 von 100 Fällen enthalten sein. Damit die Veränderung nicht rückgängig gemacht werden kann, muss sie unvorhersagbar – zufällig – sein. Bspw. reicht es nicht aus, Größenangaben einer Person um den Festwert von 5 cm zu erhöhen, da die ursprünglichen Werte durch Subtraktion von 5 cm errechenbar wären. Würde statt eines festen Wertes ein zufälliger Wert aus dem Bereich -15 bis +15 cm gewählt, wären die ursprünglichen Werte nicht errechenbar.

Kritisch für den Erfolg des Verfahrens ist die Auswahl, wie die Werte verändert werden **(„Störgrößen“)**. Führt die Wahl der Störgrößen zu übertriebenen Ergebnissen, bspw. Körpergrößen von 250 cm für Menschen, lassen sich einerseits die veränderten Merkmale bestimmen und die Veränderung – zumindest auf der Basis von Annahmen – herausrechnen. Bspw. kann man annehmen, dass die 250 cm zu einer sehr großen Person gehören. Wenn man annimmt, dass ein sehr großer Mensch 210 cm groß ist, würde die Störgröße 40 cm betragen.

Werden durch die stochastische Überlagerung logische Zusammenhänge zwischen Merkmalen eines Datensatzes aufgehoben, kann ein Angreifer dieses Wissen ebenfalls zur Rekonstruktion des Personenbezugs nutzen.

Die Veränderung führt zu einem **Informationsverlust**. Dessen Höhe lässt sich als Quotient aus der maximal möglichen Veränderung – hier 15 cm – und dem maximalen Grundwert – bspw. 210 cm – errechnen. In dem Beispiel betrüge der Informationsverlust 15 cm/210cm = 0,07 = 7 %. Ob der Informationsverlust für eine hinreichende Anonymisierung ausreicht, ist im Einzelfall zu beurteilen.

Die Anwendung der stochastischen Überlagerung reicht häufig für eine Anonymisierung nicht aus, d.h. sie muss durch **weitere Techniken** ergänzt werden.

6.3.1.1.2 Vertauschung

Die **Vertauschung** lässt die Werte der Merkmale unangetastet. Vielmehr werden Werte zwischen den Datensätzen vertauscht. Damit eignet sich das Verfahren sowohl für **qualitative Daten** (Rating, Listen) wie auch **quantitative Daten**. Bspw. erfolgt ein Vertauschen der Körpergröße zwischen Datensatz „154“ und Datensatz „357“. Voraussetzung ist, dass die Originaldaten nach Anwendung gelöscht werden.

Durch das Vertauschen wird die **Zuordnung** zwischen **Merkmal und Datensatz** entfernt.
Wird nur ein Datenfeld aus einer logischen Beziehung oder statistischen Korrelation getauscht, lässt sich der Personenbezug wiederherstellen. Dazu ist lediglich notwendig zu wissen, welche vertauschten Merkmale in einer Beziehung zu nicht vertauschten Merkmalen stehen. Bspw. wird das Gehalt zwischen „154“ und „357“ vertauscht. Die Position der Person wird nicht getauscht. Weiterhin hebt eine zufällige Vertauschung nicht zwangsläufig starke **statistische Zusammenhänge** zwischen Merkmalen auf. Deshalb ist nach der Vertauschung zu prüfen, ob der Personenbezug durch das Ausnutzen solcher Beziehungen wieder hergestellt werden kann.

Erfolgt das Vertauschen für jedes Merkmal getrennt, können die (statistischen) Eigenschaften des Datensatzes verändert werden. Deshalb müssen die Merkmale, deren (statistischer) Zusammenhang erhalten bleiben soll, zwischen den gleichen Datensätzen getauscht werden. Bspw. wird das Gehalt und die Position zwischen den Datenätzen „154“ und „357“ vertauscht. Da die Werte der Daten unverändert bleiben, ändert sich der Informationsgehalt nicht.

Nicht jede Vertauschung führt automatisch zu einer Anonymisierung. Es ist darauf zu achten, dass die Merkmale vertauscht werden, die ursächlich für den Personenbezug sind.

6.3.1.1.3 Differential Privacy

Differential Privacy[19] ändert die Originaldaten nicht, so dass diese nach Anwendung des Verfahrens nicht gelöscht werden müssen. Der Originaldatenbestand bleibt unverändert und personenbezogen. Voraussetzung ist, dass die Daten **quantitativ** sind.

Das Konzept der Differential Privacy stellt verschiedenen Nutzern jeweils eine **eingeschränkte Sicht** auf den Originaldatenbestand bereit. Dabei wird die Anzahl der Datensätze eingeschränkt. Für die angezeigten Datensätze werden die Werte der Merk-

[19] S. hierzu auch Ostendorff, OpenRedact Anonymisierungsleitfaden Open Data und Datenschutz, abrufbar unter https://openredact.org/leitfaden-anonymisierung (letzter Zugriff am 28.11.2022).

male in der Anzeige mittels mathematischer Funktionen verändert. Damit sieht der Nutzer andere Daten als im Originaldatensatz stehen. Welcher Art die Veränderung ist und wie diese technisch erfolgt, hängt vom Einzelfall ab.

Das Konzept der Differential Privacy liefert lediglich einen mathematischen Rahmen und ein Vorgehen, um die Veränderung zu bestimmen. Die Veränderung wird für jede Abfrage neu berechnet.

Zwar werden bei jeder Abfrage die gleichen Daten immer anders verändert. Gleichwohl erlaubt die Kombination mehrfacher Abfragen, die mathematische Veränderung der Werte zu ermitteln und „herauszurechnen“. Aus diesem Grund muss sichergestellt werden, dass mehrfache Abfragen einer Stelle unterbunden werden. Bei der Beurteilung, ob das Ergebnis anonym ist, ist auch der Zugriff auf die angezeigten Daten zu berücksichtigen.

6.3.1.2 Verfahren der Generalisierung

Bei einer **Generalisierung** werden Werte in ihrer Größenordnung „vergröbert“. Bspw. wird ein Straßenname durch die Postleizahl ersetzt. Durch dieses „Vergröbern“ soll das Herausgreifen von Personen erschwert werden. Die übrigen Risiken der Verknüpfbarkeit und der Inferenz ändern sich grundsätzlich nicht. Um die statistischen Eigenschaften möglichst zu erhalten, erfolgt das „Vergröbern“ regelhaft.

6.3.1.2.1 Aggregation und k-Anonymität

Bei der **Aggregation** und **k-Anonymität** werden Datensätze zu Gruppen zusammengefasst. Datensätze in einer Gruppe erhalten den gleichen Merkmalswert. Bspw. werden Gehaltsdaten durch Intervalle „20-30.000“ und „30-40.000“ ersetzt.

Wenn die Werte des zu aggregierenden Merkmals nicht gleich verteilt sind, sondern manche Werte deutlich seltener oder häufiger vorkommen als andere, besteht das Risiko, dass es Gruppen gibt, die nur einen Datensatz enthalten. Um das zu verhindern, legt die k-Anonymität fest, dass in jeder Gruppe mindestens „k“ enthalten sein müssen. Der Parameter k beschreibt die Mindestgröße einer Gruppe. Zwischen Gruppengröße k und dem Informationsgehalt besteht ein Zielkonflikt. Je Größer k ist, desto kleiner ist der Informationsgehalt. Bspw. wäre das Intervall „1-100 Mio.“ als Gehaltsangabe wenig aussagekräftig.

Häufig können **mehrere Merkmale** zu einer Identifikation der Person führen. Deshalb sind alle Merkmale, die für eine Identifikation der Person geeignet sein könnten („mittelbare Identifikatoren“), in die Gruppenbildung einzubeziehen. Eine Gruppe besteht dann aus mehreren Merkmalen. Andernfalls können die unverfälschten Werte der mittelbaren Identifikatoren genutzt werden, um Personen aus einer Gruppe zu identifizieren. Werden bspw. etwa im Rahmen einer medizinischen Studie die Klarnamen der Probandinnen und Probanden durch eine Zeichenfolge ersetzt, daneben aber die Krankheitsdiagnose, Postleitzahl und das Geburtsdatum angegeben, so kann zur Schaffung von k-Anonymität das Geburtsdatum durch das Geburtsjahr ersetzt werden.

Befinden sich in den Datensätzen drei Personen, deren Einträge bezüglich bestimmter Merkmale übereinstimmen, z.B. gleiches Geburtsjahr oder gleiche Postleitzahl, so liegt die k-Anonymität bei 3. Eine (eindeutige Zuordnung) der Krankheitsdiagnose zu einer spezifischen Person ist nicht mehr möglich.

In der Praxis besteht die Herausforderung, alle mittelbaren Identifikationen zu berücksichtigen, d.h. diese müssen vorher ermittelt werden. Besonderes Augenmerk ist auch auf Merkmale zu legen, die einen hohen Informationsgehalt aufweisen, bspw. weil sie sehr selten vorkommende Werte haben.

Ein zu kleiner k-Wert stellt ebenfalls die Anonymisierung in Frage. Bei der Gruppenbildung ist auch darauf zu achten, dass einzelne Datensätze kein zu großes Gewicht erhalten. Dieses Risiko besteht besonders bei einer ungleichmäßigen Verteilung der Werte von Merkmalen.

6.3.1.2.2 l-Diversität und t-Closeness

Das Konzept der **l-Diversität** entwickelt das Konzept der k-Anonymität weiter. Die k- Anonymität legt die Mindestgruppengröße k fest. Wie häufig ein einzelner Wert in der Gruppe vorkommt, ist nicht festgelegt.

> **Beispiel:** Von Beschäftigten eines Unternehmens werden Alter und Gehalt gespeichert. Auf das Gehalt wird das Konzept der k-Anonymität angewendet. In der Gruppe „20-30.000 Euro“ sind die Personen A (30 Jahre), B (35 Jahre) und in der Gruppe „30-40.000 Euro“ die Personen C (65 Jahre) und D (43 Jahre). Kombiniert ein Angreifer weiteres Wissen mit dem Datensatz, erkennt er, dass nur ein 65 Jähriger im Unternehmen tätig ist. Damit folgert der Angreifer, dass der 65 Jährige 30-40.000 Euro verdient.

Der Personenbezug ließe sich in dem Beispiel herstellen, weil der Wert 65 Jahre einmal in der Gruppe vorkam. Die l-Diversität verlangt, dass in jeder der „k“ Gruppen jedes Merkmal mindestens „l“ verschiedene Werte hat. Statt einem 65 Jährigen hätten „l“ 65-jährige in der Gruppe sein müssen. Voraussetzung zur Anwendung der l-Diversität ist, dass im Datenbestand **hinreichend viele Datensätze** die **gleichen Werte** besitzen.

Nimmt man als dritte Anforderung, dass die Verteilung der Werte in einer Klasse der Verteilung dieser Werte in den Originaldaten entsprechen soll, spricht man **von t-Closeness**. Voraussetzung ist, dass der Datenbestand hinreichend viele und geeignete Werte aufweist.

Zusammengenommen sind die Datensätze so in Gruppen einzuteilen, dass mindestens „k“ Datensätze in einer Gruppe sind, jeder Wert mindestens „l“ mal vorkommt und dass jeder Wert in der Gruppe so oft vorkommt, wie er im Originaldatensatz über alle Datensätze betrachtet vorkommt.

6.3.1.3 Synthetische Daten

Die bei der Verarbeitung personenbezogener Daten zu beachtenden datenschutzrechtlichen Vorgaben können das Ausmaß und die Art der Datenverarbeitung einschränken. Eine mögliche Alternative zur Nutzung personenbezogener Daten ist für bestimmte Anwendungsfälle die Nutzung von **synthetischen Daten**. Im Gegensatz zu personenbezogenen Daten werden synthetische Daten nicht zu bestimmten natürlichen Personen erfasst. Dementsprechend stellen sie auch keine Information über natürliche Personen zur Verfügung. Vielmehr handelt es sich bei synthetischen Daten um durch ein **Berechnungsverfahren erzeugte Daten**.

Zur Erzeugung synthetischer Daten werden **Synthesemodelle** zugrunde gelegt. Diese bestimmen wesentliche Eigenschaften des Berechnungsverfahrens, mit dem die synthetischen Daten erzeugt werden. Zu den bekannten Synthesemodellen zählen

- Modelle, die Daten zufällig aus einer Liste vorgegebener Beispiele auswählen. Ein Beispiel ist die zufällige Auswahl von Städten aus einer Liste von nicht bestimmten Betroffenen zuordenbaren Einträgen.
- Modelle, die synthetische Daten aus der zufälligen Aneinanderkettung von Zeichenfolgen aus einem Alphabet erzeugen
- Modelle, die Daten nach vom Menschen fest vorgegebenen Regeln erzeugen. Ein Beispiel ist die Regel, Daten in einem bestimmten Format zu erzeugen oder die Regel, Daten aus einer Liste von weiblichen Vornamen französischen Ursprungs zu wählen.
- Modelle, die Daten nach aus einer KI ermittelten Regeln erzeugen. Dies können Zusammenhänge sein, die der Ersteller des Berechnungsverfahrens ohne Hilfe der KI nicht erkannt hätte.

Synthetische Daten, die aus mehreren Datenfeldern bestehen, können auch aus einer Kombination der drei Arten erzeugt werden.

Synthetische Daten können personenbezogenen Daten in bestimmten Eigenschaften **nachempfunden sein**. Diese Eigenschaften werden durch das Berechnungsverfahren so nachgebildet, dass die berechneten synthetischen Daten anstelle der personenbezogenen Daten nutzbar sind. Die Nutzbarkeit macht es erforderlich, dass die synthetischen Daten den echten personenbezogenen Daten ähneln. Der Grad der Ähnlichkeit wird durch den Anwendungsfall bestimmt. Gleichzeitig muss beachtet werden, dass die synthetischen Daten keine Re-Identifizierung Betroffener zulassen.

Ein Beispiel für den Einsatz synthetischer Daten sind Daten, die zum Testen der Funktionalität neu entwickelter, mit personenbezogenen Daten zu nutzender Software verwendet werden sollen. Anstatt Software mit echten personenbezogenen Daten zu testen, kommen synthetische Daten zum Einsatz. Um aussagekräftige Testergebnisse zu erhalten, müssen diese den echten Daten **ausreichend ähneln**. Die Ähnlichkeit wird u.a. erreicht, indem bei der Erzeugung der synthetischen Daten **statistische Eigenschaften** der personenbezogenen Daten durch das Berechnungsverfahren nachgebildet werden. Hierbei ist sicherzustellen, dass die synthetischen Daten den personenbezogenen Daten nicht derart ähneln, dass sie zur Re-Identifizierung Betroffener führen können. Hierfür muss das Berechnungsverfahren sorgfältig entworfen werden. Neben

dem Erhalt bestimmter Eigenschaften der zugrundeliegenden personenbezogenen Daten muss beim Entwurf des Berechnungsverfahrens beachtet werden, dass die Nachbildung der Eigenschaften nicht zu synthetisierten Daten führt, die eine Re-Identifizierung Betroffener ermöglicht. Ein Beispiel ist die Nachbildung eines Datensatzes aus Vor- und Nachnamen weiblicher Personen aus dem deutschen Sprachraum. Bei der Übernahme von Wortpaaren (Vorname, Nachname) ohne Änderung aus dem zugrundeliegenden echten Datensatz in den synthetischen Datensatz können insbesondere selten vorkommende Namenskombinationen zur Re-Identifizierung Betroffener genutzt werden.

6.4 Bewertungsmatrix

Technik	Anwendbar auf	Herausgreifen	Verknüpfbarkeit	Inferenz
Stochastische Überlagerung	1	-	-	o
Vertauschung	1,2	o	-	-
Differential Privacy	1	+	- (mehrfach Anwendung)	- (mehrfach Anwendung)
Aggregation und k-Anonymität	1,2	+	o	-
L-Diversität und t-Closeness	1,2	+	o	o
Synthetische Daten	1,2	+	+	+

Tabelle 1: Übersicht zur Wirkung von Anonymisierungstechniken (1 = quantitative Daten, 2 = qualitative Daten, - weiterhin möglich, o möglich aber erschwert, + verhindert)[20]

6.5 Abgrenzung zu anderen Verfahren

6.5.1 Hashfunktion

Eine **Hashfunktion** ist ein mathematisches Verfahren, dass im Prinzip eine Zeichenkette durch eine kürzere Zeichenkette ersetzt. Aus einem Namen, wie z.B. „Musterbetroffener“, wird „AF341“.

Sofern ein Angreifer weiß, welcher Algorithmus, zum Berechnen der Hashwerte verwendet wurde, kann er die Hashwerte der für ihn interessanten Werte berechnen und seinen berechneten Hashwert mit dem „anonymen“ Hashwert vergleichen. Stimmen

[20] Die Darstellung basiert im Wesentlichen auf Art- 29 Gruppe (2014): Stellungnahme 5/2014 zu Anonymisierungstechniken. Angenommen am 10. April 2014 (WP 216).

beide Hashwerte überein, hat der das Originaldatum, in dem Beispiel den Namen, wahrscheinlich ermittelt. Es ist keine sichere Ermittlung, da ein „Hashen“ verschiedener Originalwerte zum gleichen Hashwert führen kann. Um dieses „Zurückrechnen“ zu erschweren, können bei kryptographischen Hashfunktionen weitere zufällige Werte hinzugefügt werden (**„Salt“** und **„Pepper“** genannt). Dadurch würde beim „Hashen“ von „Musterbetroffenen“ immer ein anderer Hashwert erzeugt. Erfährt der **Angreifer**, wie er Pepper und Salt berechnen kann, kann er eventuell ebenfalls die eine Re-Identifizierung durchführen. Auf jeden Fall werden die statischen Eigenschaften des gehashten Merkmals zerstört.

Mittels Hashfunktionen lassen sich Daten nicht anonymisieren.[21] Das Ergebnis des Hashvorganges ist bestenfalls Pseudonymität.

7. Einbeziehung von Dritten oder Auftragsverarbeitern

7.1 Weitergabe an Dritte

Es ist möglich, dass anonymisierte Daten an einen **Dritten** und damit eigenständigen Verantwortlichen **weitergegeben werden**. Werden die Daten vor der Übermittlung hinreichend anonymisiert, sind die datenschutzrechtlichen Anforderungen der DS-GVO allgemein (so z.B. die Zulässigkeit der Weitergabe oder deren Transparenz gegenüber Betroffenen) nicht zu beachten. Der Dritte als Empfänger der Daten wird jedoch zu überprüfen haben, ob die Daten **für ihn** unter Berücksichtigung der bei ihm verfügbaren Mittel und der Wahrscheinlichkeit ihres Einsatzes anonym sind. Ein **vertragliches Verbot** einer Re-Identifizierung vermag keine wirksame Maßnahme sein, um eine solche per se auszuschließen. Sind die weitergegebenen Daten für ihn nicht anonym, fällt die Verarbeitung erneut in den **Anwendungsbereich** der DS-GVO.

Eine gesetzliche Pflicht zur Einbindung eines Dritten, z.B. als **Vertrauensstelle**, besteht bei der Anonymisierung im Übrigen nicht.

7.2 Gemeinsame Verantwortlichkeit

Entscheiden zwei oder mehr Verantwortliche **gemeinsam über Zwecke und Mittel** der Verarbeitung personenbezogener Daten und bildet die Anonymisierung einen Teil der Verarbeitungsvorgänge, sind die Anforderungen aus Art. 26 DS-GVO an die Vereinbarung zwischen den Beteiligten zu beachten. Dort sollten die **Rahmenbedingungen der Anonymisierung** beschrieben werden. Führt ein Verantwortlicher im Rahmen der gemeinsamen Verantwortlichkeit die Anonymisierung durch, sollte er in der Vereinbarung als Verantwortlicher für das Verfahren der Anonymisierung benannt werden. Gewährt dieser Verantwortliche einen Zugriff auf anonymisierte Daten, muss über ein **Rechte- und Rollenkonzept** sichergestellt sein, dass ein Zugriff der übrigen Verantwortlichen auf einen Originaldatensatz ausgeschlossen ist.

[21] Anderer Ansicht: BDI (2020): Anonymisierung personenbezogener Daten, S. 21.

Ferner bietet sich die **vertragliche Verpflichtung** der Beteiligten hinsichtlich der **Prüfung** an, ob sich aus einem vorhandenen eigenen Datenbestand durch Abgleich mit einem bspw. übermittelten anonymisierten Datensatz eine natürliche Person bestimmen lässt. Für den Fall einer erkannten Re-Identifizierbarkeit sollten die Folgen in der Vereinbarung geregelt werden. So z.B. das Aufleben der Anforderungen der DS-GVO hinsichtlich einer Verarbeitung personenbezogener Daten.

7.3 Auftragsverarbeiter

Die DS-GVO kennt zwei Rollen, die Unternehmen einnehmen können: **Verantwortlicher** (engl. „Controller“) und **Auftragsverarbeiter** (engl. „Processor“). Beispielsweise kann ein Dienstleister Lohn- und Gehaltsabrechnungen für Arbeitgeber durchführen. Für diese Abrechnung im Auftrag agiert der Dienstleister als Auftragsverarbeiter. Für den eigenen Recruiting-Prozess ist er jedoch Verantwortlicher. Ein Auftragsverarbeiter agiert als Erfüllungsgehilfe für seinen Auftraggeber. Damit ist ein Auftragsverarbeiter ein **weisungsgebundener Dienstleister**. Ein Verantwortlicher bestimmt über Zwecke oder Mittel der Verarbeitung. Er ist „Herr der Daten“.

7.3.1 Der Auftragsverarbeiter anonymisiert für den Verantwortlichen

Werden personenbezogene Daten weisungsgebunden durch einen Auftragsverarbeiter gem. Art. 4 Nr. 8 DS-GVO anonymisiert (z.B. durch ein durch ihn gehostetes Anonymisierungstool), sind die Anforderungen aus Art. 28 DS-GVO zu beachten. Immerhin werden hier personenbezogene Daten verarbeitet (s. auch Ziff. 4.2). Abzugrenzen ist dieser Fall von einem Anonymisierungstool, das der Verantwortliche selbst betreibt.

Fraglich ist jedoch, ob mit Blick auf die **Weisungsgebundenheit** des Dienstleisters die Auftragsverarbeitung ein **geeignetes Mittel** für die Anonymisierung darstellt. Immerhin könnte der Verantwortliche theoretisch per Weisung den Dienstleister zur **Offenlegung der verwendeten Anonymisierungstechni**k verpflichten. Darüber hinaus verfügt eine weitere Stelle ggf. über das **Wissen** um die durchgeführte Anonymisierungstechnik, was sich ein Angreifer zunutze machen könnte. Möchte ein Verantwortlicher personenbezogene Daten durch einen Auftragsverarbeiter anonymisieren lassen, müssen zumindest **vertragliche Regelungen** getroffen werden, die eine Offenlegung der Anonymisierungstechnik und deren Durchführungsschritte verbieten. Vorteilhafter ist es, eine Software zur Anonymisierung selbst zu betreiben

7.3.2 Der Auftragsverarbeiter anonymisiert für eigene Zwecke

7.3.2.1 Zulässigkeit der Verarbeitung

Was passiert, wenn der weisungsgebundene Dienstleister Daten für eigene Zwecke nutzen und sie hierfür vorher anonymisieren möchte?

Die Anonymisierung ist eine Datenverarbeitung, für die es einen Erlaubnistatbestand in der DS-GVO geben muss.

Sobald ein Auftragsverarbeiter Daten des Auftraggebers für eigene Zwecke nutzt - und sei es „nur“ durch eine Anonymisierung - **wird er für diese Verarbeitung Verantwortlicher** (Art. 28 Abs. 10 DS-GVO). Der weisungsgebundene Dienstleister schwingt sich zum „Herren der Daten“.

Der Verarbeitung des Dienstleisters für eigene Zwecke bedeutet mit Blick auf die Rechtmäßigkeit der Verarbeitung, dass

- der ursprünglich Verantwortliche die **Kompatibilität der Zweckänderung** zu prüfen hat (Art. 6 Abs. 4 DS-GVO) und
- der Dienstleister eine eigene **Rechtsgrundlage** für die Datenverarbeitung benötigt.

Ob der Dienstleister eine **Rechtsgrundlage** für eine Verarbeitung zu eigenen Zwecken hat, ist sorgfältig zu prüfen. Dies gilt auch für den Verarbeitungsvorgang der Anonymisierung. Immerhin agiert er hinsichtlich der Datenverarbeitung eng **vertrags- und weisungsgebunden**.

Teilweise lassen Auftragsverarbeiter sich vertraglich Zugriffs- und Verwertungsrechte an den Daten des Auftraggebers allgemein einräumen. Diese Rechteeinräumung allein legitimiert die damit einhergehende Datenübermittlung nicht. Die Einräumung eines Rechts zur eigennützigen Datenverarbeitung durch den Dienstleister muss mit der ursprünglichen Rechtsbeziehung des Verantwortlichen zum Betroffenen **kompatibel** sein. Eine Einwilligung des Betroffenen für die Datenverarbeitung und damit auch für die Anonymisierung wird in der Regel nicht vorliegen.

7.3.2.2 Sanktionen

Wird ein Auftragsverarbeiter beauftragt, der erkennbar Daten für eigene Zwecke auf inkompatible Art und Weise verarbeiten will, läuft der Auftraggeber Gefahr, gegen das Gebot zur **sorgfältigen Auswahl** des Dienstleisters zu verstoßen (Art. 28 Abs. 1 DS-GVO). Dieser Verstoß kann gemäß Art. 83 DS-GVO mit einem Bußgeld von bis zu 10 Millionen Euro oder bis zu 2 % des weltweiten Jahresumsatzes sanktioniert werden.

Durch die weisungswidrige Datenverarbeitung für **eigene Zwecke** wird der Auftragsverarbeiter selbst zum Verantwortlichen (Art. 28 Abs. 10 DS-GVO). Hat er keine Rechtsgrundlage für diese Datenverarbeitung, ist diese **rechtwidrig**. Eine rechtswidrige Datenverarbeitung kann gemäß Art.83 DS-GVO mit einem Bußgeld von bis zu 20 Millionen Euro oder bis zu 4 % des weltweiten Jahresumsatzes sanktioniert werden.

Unter Umständen, z.B. bei einer fehlgeschlagenen oder unzureichenden Anonymisierung, kommt auch gemäß Art. 82 DS-GVO ein Ersatz des **materiellen oder immateriellen Schadens** des Betroffenen in Betracht, für den der Verantwortliche und der Dienstleister **gesamtschuldnerisch** haften.

7.4 Anonymisierung innerhalb der Unternehmensgruppe

Wird die Anonymisierung innerhalb einer **Unternehmensgruppe** durchgeführt oder werden anonymisierte Daten weitergegeben, sind die allgemeinen Anforderungen an die Einbeziehung von Dritten zu beachten. Insoweit kennt die DS-GVO kein echtes **Konzernprivileg**. Es wird bei der Prüfung einer hinreichenden Anonymisierung darauf ankommen, welche Mittel bspw. einem Mutterkonzern zur Verfügung stehen, um ein eine natürliche Person zu identifizieren bzw. wie wahrscheinlich deren Einsatz ist. Solche Mittel können grundsätzlich auch eine Weisung sein, die ein Mutterkonzern gegenüber einem Tochterunternehmen hinsichtlich der Herausgabe von Daten ausspricht.

8. Ausgewählte Einsatzklassen

In Kapitel 6.3.1 wurden ausgewählte Verfahren zur Anonymisierung personenbezogener Daten vorgestellt. Ziel der Anwendung von Anonymisierungsverfahren auf personenbezogene Daten ist die Generierung von Daten, die für bestimmte Anwendungsfälle nutzbar sind, jedoch keinen Personenbezug mehr aufweisen. Dies ist der Fall, wenn die generierten Daten ohne unverhältnismäßigen Aufwand nicht mehr zur Re-Identifizierung Betroffener genutzt werden können. Die Eigenschaften anonymisierter Daten unterscheiden sich je nach Anwendungsfall.

Im vorliegenden Kapitel werden die ausgewählten Einsatzklassen **‘Anonymisierung als Löschung’**, **‘Anonymisierung bei Weitergabe’**, **‘Anonymisierung zum Training von Algorithmen’** und **‘Anonymisierung zum Testen von Software’** für die Nutzung von anonymisierten Daten vorgestellt. Ziel ist in erster Linie die exemplarische Veranschaulichung der Möglichkeit, aus personenbezogenen Daten für den Anwendungsfall in der Praxis passgenaue Anonymisierungen bereitzustellen. Weiterhin sollen mögliche Grenzen der Nutzbarkeit anonymisierter Daten aufgezeigt werden.

Die Anonymisierung personenbezogener Daten hat u.a. zum Ziel, das Risiko einer unerlaubten Kenntnisnahme persönlicher Umstände bzw. die Re-Identifizierung Betroffener durch Unbefugte, sogenannte Angreifer, zu minimieren. Daher wird beim Entwurf der Anonymisierungslösung eine mögliche Vorgehensweise eines Angreifers berücksichtigt (s. hierzu auch Kapitel 6.2). Um dies zu veranschaulichen, umfasst die Struktur der im Folgenden vorgestellten Beispiele eine kurze Beschreibung des möglichen **Angreifers** und die betrachteten zu schützenden **personenbezogenen Daten**. Im Beispiel wird zunächst die gewünschte **Nutzung bzw. Verarbeitung** der personenbezogenen Daten beschrieben. Dann wird die **Anonymisierungslösung** für das Beispiel vorgestellt. Daraufhin werden die **Eigenschaften der anonymisierten Daten** beschrieben und im Sinne der Verarbeitbarkeit für die geplante Anwendung bewertet.

8.1 Anonymisierung als Löschung

Die Löschung personenbezogener Daten nach Zweckverbrauch gehört zu den Pflichten des Verantwortlichen. Betrachtet man hierbei die Löschung als das Entfernen des Personenbezugs, so kann die Anonymisierung als Löschung betrachtet werden. Hierbei

muss beachtet werden, dass die personenbezogenen Daten unmittelbar nach Erzeugung der Anonymisierung und Erlöschen der rechtlichen Grundlage der Datenverarbeitung vom datenverarbeitenden System entfernt werden müssen. Die verbleibenden anonymisierten Daten weisen keinen Personenbezug mehr auf. Ihre Nutzung ist entsprechend nicht an datenschutzrechtliche Bestimmungen gebunden.

Ein Vorteil gegenüber dem vollständigen Entfernen der Daten vom datenverarbeitenden System ist die Möglichkeit, die anonymisierten Daten unabhängig vom Verarbeitungszweck der ursprünglichen personenbezogenen Daten nutzen zu können. Die im Folgenden beschriebenen Beispiele sollen diese Möglichkeit aufzeigen. Darüber hinaus sollen Grenzen der Nutzbarkeit der als Löschung anonymisierten Daten exemplarisch aufgezeigt werden.

8.1.1 Beispiel: Eckdaten zu Bewerbungen behalten

Bewerbungen sind bekanntlich, nachdem die Einstellungsentscheidung getroffen wurde und die Verjährungsfrist von Ansprüchen, insbesondere aus dem AGG, abgelaufen ist, zu löschen. Die Löschfrist beträgt somit wenige Monate. Personalverantwortliche stehen vor der Aufgabe, den Vermittlungserfolg sowie die Qualität der Bewerber über Jahre hinweg messen zu müssen. Würden die Bewerbungsdaten vollständig gelöscht, ließen sich solche Messungen nicht oder nur für sehr kurze Zeiträume vornehmen.

Anstelle der vollständigen Löschung des Datensatzes bietet sich eine Anonymisierung an. In der Praxis wird Anonymisierung mit dem Löschen offensichtlich identifizierender Merkmale wie z.B. Name, Anschrift oder Kontaktdaten gleichgesetzt. Eine Kombination von „Skills“, bspw. in Verbindung mit Stationen im Lebenslauf, kann eine Re-Identifizierung unter Zuhilfenahme von Profilen in Sozialen Netzwerken ermöglichen. Deshalb sollte geprüft werden, ob nach dem Löschen der offensichtlich identifizierenden Merkmale weitere Anonymisierungstechniken angewendet werden müssen.

Ein **Angreifer** kann an sensiblen Informationen des Bewerbers wie Werdegang, Angaben zum Wunschgehalt und einer Schwerbehinderung interessiert sein.

Die **betrachteten personenbezogenen Daten** umfassen das Geburtsdatum, den Geburtsort, die schulische Laufbahn, Sprachkenntnisse, Staatsangehörigkeit, die Adresse und mögliche Angaben zu einer Schwerbehinderung.

Wurden die personenbezogenen Daten derart anonymisiert, dass sie als gelöscht zu betrachten sind, ist die Zuordnung einzelner Informationen zu einzelnen Bewerbern nicht mehr ohne einen unverhältnismäßigen Aufwand möglich. Dies hat zur Folge, dass bestimmte Kennzahlen über einen Bewerber aus den anonymisierten Daten nicht mehr erfasst werden können. Möglicher Nutzen der anonymisierten Daten ergibt sich jedoch im Vergleich zur Löschung für Informationen, die aus geeignet anonymisierten Daten ermittelt werden können. **Beispiele** hierfür sind in der nachfolgenden **Tabelle 3** gelistet.

Personenbezogene Daten	Anonymisierung
Jahrgänge der Bewerber	Durchschnittswert über alle Bewerber im Jahre x für mehrere Jahre. (A)
Herkunft	Generalisierung der Herkunft nach Region (z.B. Mitteleuropa, Naher Osten, Nordamerika) (V)
Sprachkenntnisse	Durchschnittswert der Anzahl beherrschter Sprachen über alle Bewerber im Jahre x für mehrere Jahre. (A) Durchschnittswert der Anzahl der Bewerber, die Englisch auf dem Niveau C1 beherrschten; im Jahre x für mehrere Jahre. (A)
Bildungsstand	Durchschnittswerte der erreichten Schulabschlüsse im Jahre x; Anzahl der Bewerber mit Bachelorabschluss. (A)
Abschlussnoten	Durchschnittsnote, aufgeteilt nach Jahren und Abschlüssen. (A)
Wohnort	Region des Wohnortes, z.B. Rheinland statt Bonn. (V)
Schwerbehinderung	Anteil Schwerbehinderter an den Bewerbern im Jahre x. (A)

Tabelle 3: Beispiele für personenbezogene Daten und die Ersetzung dieser durch geeignete Verallgemeinerung (V) bzw. Aggregate (A).

Die Anonymisierungslösung besteht hier aus der Anwendung von Verfahren zur Berechnung von Durchschnitten bzw. Generalisierung und anschließende vollständige Entfernung der ursprünglich erfassten personenbezogenen Daten. Aus den anonymisierten Daten lassen sich Aussagen über die Gesamtheit der Bewerber in einem bestimmten Zeitraum, z.B. Jahr x ableiten. Darüber hinaus ist die Zuordnung von Informationen zu einem bestimmten Bewerber nicht mehr möglich.

8.1.2 Beispiel: Qualitätsanalyse des Kundendienstes eines Elektrohändlers

Ein **Elektrohändler** bietet nach alter Tradition den Verkauf, die Wartung und die Reparatur von elektronischen Geräten an. Der **Kundendienst des Elektrohändlers** möchte wissen, welche Bedürfnisse Kunden aufweisen, die ihn besonders häufig wegen Problemen mit den zuvor erworbenen Geräten kontaktieren. Hierfür möchte er zunächst alle über einen Anruf eines Kunden anfallenden Daten speichern und diese später analysieren: Das transkribierte Kundengespräch, das zuvor gekaufte Gerät, Name, Alter und,

abgeleitet aus den Gesprächen, Technikaffinität und Bildungsstand des Kunden, Häufigkeit der Nutzung des Geräts usw.
Um die Daten der Kunden rechtssicher analysieren zu können, holt der Elektrohändler deren Einwilligung zur Analyse der Gespräche ein. Hierbei fällt ihm auf, dass 85% der Kunden diese Einwilligung nicht erteilen. Daher ist er verpflichtet, zumindest die personenbezogenen Daten dieser Kunden nach Ende des kundendienstlichen Vorgangs, der mit dem Gespräch verbunden ist, zu löschen. Der Startpunkt des kundendienstlichen Vorgangs wird in einem Ticketing-System durch ein geöffnetes Ticket markiert. Das Schließen des Tickets durch einen Mitarbeiter des Kundendienstes bedeutet die Beendigung des Vorgangs.

Lösung: Nach dem Schließen des mit dem Gespräch verbundenen Tickets reichert der Kundendienst Aggregate der zuvor erfassten personenbezogenen Daten an. Die restlichen Daten werden gelöscht.

Ein **Angreifer** kann Interesse an Details aus den Kundengesprächen haben, aus denen eine Motivation zum Wechsel zu einem anderen Händler bzw. Kundendienst abgeleitet werden kann. Auch kann er z.B. ein Interesse an der Anzahl der Aufträge und der Anzahl der Anfragen an den Kundendienst haben.

Die **betrachteten personenbezogenen Daten** umfassen

- Zeitpunkt und Dauer des Gesprächs
- Der Text des transkribierten Gesprächs
- Name des Kunden
- Alter des Kunden
- Indikatoren für Technikaffinität und Bildungsstand des Kunden, erfasst z.B. durch persönliche Angabe des höchsten Abschlusses oder durch die korrekte Nutzung von Fachbegriffen

Weitere Informationen, wie z.B. das gekaufte Gerät, können ebenfalls als personenbezogen betrachtet werden. Dies unterliegt der Einschätzung eines für den Elektrohändler zuständigen Datenschutzbeauftragten.

Die **Nutzung der Daten** umfasst zum Zeitpunkt der Öffnung des Tickets Folgendes:

- Ermittlung, welche Abteilung des Kundendienstes mit der Lösung der Kundenanfrage befasst werden soll und die Weiterleitung von Information an diese
- Rückruf und Terminvereinbarung
- Erhalt und Reparatur defekter Geräte; Zuordnung zu einzelnen Kunden; Rückgabe des reparierten Gerätes an den Kunden
- Erstattung oder teilweise Erstattung von Kaufpreisen

Nach dem Schließen des Tickets ist die Nutzung der Daten eingeschränkt auf jene Analysen bzw. Verarbeitungen, die auf aggregierten Daten ohne direkten Personenbezug basieren.

Die Daten werden bereits vor dem Schließen des Tickets anonymisiert. Dies erfolgt durch Anreicherung von Aggregatswerten durch Ergänzung um die im Rahmen der Bearbeitung des Kundenvorgangs ermittelte Information. Bei Bedarf können Zeitfenster

festgelegt werden, innerhalb derer anfallende Daten aggregiert werden. So lassen sich genauere Aussagen über die Bedürfnisse der Kunden und deren Verhalten über bestimmte Zeiträume treffen. Auch ist es möglich, Wertepaare von Aggregaten zu erzeugen. Hierdurch werden z.B. zweidimensionale Kategorisierungen der aus personenbezogenen Daten abgeleiteten Information möglich. Nach Schließen des Tickets werden die ursprünglichen personenbezogenen Daten vom System entfernt.
Beispiele für anonymisierte Daten als Ersatz für gelöschte personenbezogene Daten sind in nachfolgender **Tabelle 4** gelistet:

Personenbezogene Daten	**Anonymisierung**
Zeitpunkt und Dauer des Gesprächs	Dauer: Drei Kategorien: kurze Dauer (<5 Minuten), mittlere Dauer (<10 Minuten), lange Dauer (>=10 Minuten). Zeitpunkt: Zeitfenster: vormittags, mittags, nachmittags, abends. Bei Bedarf, zweidimensionale Kategorisierung: Dauer x Zeitpunkt.
Der Text des transkribierten Gesprächs	Typ: Allgemeine Beschwerde, Verspätete Lieferung, defekte Ware, Gewährleistungsfall, Garantiefall, Reparaturanfrage, …Bei Bedarf, mehrdimensionale Kategorisierung (z.B. verspätete Lieferung, defekte Ware).
Name des Kunden	entfernt
Alter des Kunden, z.B. durch Erfassung des Geburtsdatums	Alterskategorie: 18-25, 26-35, 36-45, 46-55, 56-65,>65.
Gekauftes Gerät	Bleibt erhalten
Indikatoren für Technikaffinität und Bildungsstand des Kunden	Durchschnitt der Kunden, die Fachbegriffe korrekt nutzen. Durchschnitt der Kunden, die mindestens das Abitur erreicht haben.

Tabelle 4: Information, die ohne Erhaltung des Personenbezugs nach Löschung vorhanden sein soll.

8.1.3 Beispiel: Webseitenstatistiken

Der Betreiber einer Online-Einkaufsplattform bietet die Möglichkeit, Waren zu bestellen und zu liefern. Zur Vereinfachung von Bestellungen können Kunden Profile anlegen, über die Bestellungen vereinfacht abgewickelt werden können, ohne, dass Kunden bei

jeder erneuten Bestellung ihre Daten neu eintragen müssen. Darüber hinaus ist es möglich, gefundene Produkte zu markieren, um sie später einfacher wiederzufinden sowie Präferenzen für neue Angebote aus bestimmten Kategorien festzulegen.

Angreifer im Bereich von Webseitenstatistiken sind in diesem Beispiel im internen Bereich angesiedelt. Natürlich wäre auch der Abfluss von Kundenprofilen an externe Firmen oder Institutionen möglich, derartige Beispiele werden jedoch im Kapitel 8.2 behandelt.

In diesem Szenario besteht die Hauptgefahr einer unzulässigen, zweckfremden Nutzung der Daten. Beispielsweise werden die Daten nicht mehr rein zur Abwicklung von Bestellungen, sondern in weiteren Analysen verwendet, welche letztlich zu einer Umsatzsteigerung führen sollen. Darüber hinaus können auch im Nachhinein Nutzer identifiziert und Bestellungen zugeordnet werden.

Betrachtete personenbezogene Daten sind primär Versandinformationen wie der vollständige Name, die zugehörige Liefer- und Rechnungsadresse und die Telefonnummer zur Kontaktaufnahme im Problemfall oder bei größeren Speditionslieferungen. Benötigt werden darüber hinaus die E-Mail-Adresse zur Kontoführung sowie Zahlungsinformationen. Diese beinhalten die Art der Zahlung neben möglicher Nummern, wie der Kreditkartennummer.

Aus dem aktiven Nutzen der Plattform ergeben sich Informationen zu für die jeweilige Person interessanten Produkten oder Produktgruppen.

Die **Verarbeitung der Daten** geschieht zunächst im Rahmen des vorgesehenen und vom Kunden genehmigten Zwecks der Bestellungsabwicklung sowie Services zur Vereinfachung des Nutzens der Plattform. Widerspricht der Kunde sofort oder später einer weitergehenden Nutzung der Daten, muss diesem Rechteentzug Rechnung getragen werden.

Offen bleibt hierbei die Frage, ob die Löschung von Daten zwingend notwendig ist und welche dies betreffen sollte. Problematisch ist hierbei, dass jegliche Nutzbarkeit der Daten verloren ginge.

Anonymisierungslösungen können in diesem Fall die Nutzbarkeit der Daten größtenteils aufrechterhalten. Werden im Vorhinein Kategorien festgelegt, ist es möglich, diese zu einer Kategorisierung der Nutzer zu verwenden. Somit kann der Name entfernt und lediglich das Geschlecht gespeichert werden. Sollte das Alter angegeben sein, kann dieses generalisiert werden oder in Kategorien wie “jugendlich” oder “junger Erwachsener”, …, “Senior” eingeordnet werden. Adressen können zu (größeren) Postleitzahlengebieten generalisiert werden. Telefonnummern könnten entfernt oder analog zu Postleitzahlen generalisiert werden und Zahlungsinformationen wiederum in Kategorien eingeordnet werden. Diese würden beispielsweise “Bankeinzug”, “Nachname” oder “Sofortüberweisung” lauten.

Die **Eigenschaften der anonymisierten Daten** erfüllen weiterhin die Hauptaspekte der Nutzbarkeit, jedoch ist bei Beachtung von Gütekriterien, wie *k-Anonymität* oder weiteren, der Personenbezug als entfernt zu betrachten und die Daten unterliegen nicht mehr der DS-GVO.

Gewünschte statistische Analysen bleiben auf den anonymisierten Daten jedoch möglich. Beispielsweise lässt sich feststellen ob Personen aus gewissen geografischen Regionen oder Altersklassen gewisse Produktkategorien bevorzugen oder eher höherwertige Produkte wählen. Dies kann zur Optimierung der Werbung oder Produktanzeige verwendet werden.

In der nachfolgenden **Tabelle 5** ist beispielhaft ein vereinfachter Auszug aus den Bestellungen eines fiktiven Fahrradhändlers zu sehen, der sich auf die zum Verständnis des Vorgehens wesentlichen Merkmale konzentriert. Erfasst werden Name, Geschlecht, Adresse, Alter, E-Mail, Zahlungsart mit weiteren Informationen zur eventuellen Rücküberweisung und der Einkauf.

Im Rahmen einer **Marketingkampagne** und zur **Optimierung des Onlineshops** möchte der Versandhändler wissen, welchen Personen welche Produkte angezeigt werden sollen. Dazu erstellt er Tabelle 5, welche diese Analysen immer noch ermöglicht, aber keinen Personenbezug mehr aufweist und zudem 2-Anonymität, also k-Anonymität mit k=2, erfüllt.

Hiermit lässt sich ermitteln, dass im Gebiet der generalisierten Postleitzahl 5311* eher niedrigpreisige Fahrräder abzusetzen sind, während im Bereich 5322* hochpreisige Fahrräder gekauft werden. Diese Kunden waren ebenso männlich, während weibliche oder diverse eher etwas günstigere Produkte bevorzugten. Anhand der Kategorisierung nach Alter lässt sich feststellen, dass junge Erwachsene hochpreisige sportliche Räder bevorzugen, während Erwachsene eher normale bis günstige bevorzugen. Senioren hingegen kaufen Fahrräder, die sich durch ihre Bequemlichkeit auszeichnen.

Mit mehr Merkmalen in der Tabelle und einer größeren Datengrundlage durch mehr Tabelleneinträge lassen sich diese Analysen noch deutlich erweitern und verfeinern. Allerdings zeigt bereits dieses Beispiel, dass Marktanalysen auch auf anonymisierten Daten weiterhin möglich bleiben.

Name	**Geschlecht**	**Adresse**	**Alter**	**E-Mail**	**Zahlungsart**	**Einkauf**
Max Muster	d	Musterweg 1 53115 Bonn	40	muster@bspmail.de	Giropay DE77 8765 7896 8907 8970 00	“Drahtesel” 799€
Luise Müller	w	Waldweg 7 53115 Bonn	65	lm379@bspmail.de	PayPal DE86 9872 1234 5674 5678 32	“Hollandrad Super” 577€

Maximilian Müller	m	Hauptstraße 3 53229 Bonn	21	max@bspmail.de	Sofortüberweisung DE65 2853 4637 7531 8953 55	“Sportrad fix” 1699€
Edgar Michels	m	Lärchenweg 3 53225 Bonn	19	em@bspmail.de	PayPal DE53 5674 1842 5953 0000 00	“Mountainbike Rodeo” 2099€
Paris Gebhart	d	Wiesenweg 3 53117 Bonn	38	paris777@bspmail.de	PayPal DE22 0056 7431 8642 4533 33	“Budgetbike” 299€
Marianne Henschel	w	Musterweg 12 53115 Bonn	66	henschel@bspmail.de	Nachnahme	“Rad bequem” 649€

Tabelle 5: Datenbank eines Fahrradhändlers

Name	**Geschlecht**	**Adresse**	**Alter**	**E-Mail**	**Zahlungsart**	**Einkauf**
-	d	5311*	Erwachsener	-	Giropay	“Drahtesel 7” 799€
-	w	5311*	Senior	-	PayPal	“Hollandrad Super” 577€
-	m	5322*	Junger Erwachsener	-	Sofortüberweisung	“Sportrad fix” 1699€
-	m	5322*	Junger Erwachsener	-	PayPal	“Mountainbike Rodeo” 2099€

-	d	5311*	Erwachsener	-	PayPal	“Budgetbike” 299€
-	w	5311*	Senior	-	Nachnahme	“Rad bequem” 649€

Tabelle 6: Anonymisierte Datenbank eines Fahrradhändlers

8.2 Weitergabe anonymisierter Daten

Wenn Daten an Dritte weitergegeben werden, dürfen - je nach Verarbeitungszweck - keine personenbezogenen Daten mehr vorhanden sein, die eine Identifizierung bestimmter Personen ermöglichen. Die hier vorgestellten Beispiele stellen Situationen vor, in denen Daten weitergegeben werden und machen Vorschläge über die dabei anwendbaren Techniken, durch die eine gewünschte Verarbeitung oder Nutzung weiterhin möglich ist, ohne einen eindeutigen Personenbezug zu besitzen.

8.2.1 Weitergabe von Gehaltslisten

Die Gehaltsstruktur verschiedener Unternehmen aus der gleichen Branche hilft Personalverantwortlichen, Gehaltswünsche von Mitarbeitern einzuordnen und die Attraktivität der Gehaltsstruktur im Markt einzuschätzen. Anbieter von Gehaltsvergleichen erfassen die Gehaltsdaten direkt bei interessierten Personen oder fragen vollständige Gehaltslisten mit Merkmalen wie Qualifikation, Position, Branche, Arbeitsort, Alter, Gehaltsbestandteile, Geschlecht direkt bei Unternehmen ab.

Für eine Weitergabe personenbezogener Gehaltslisten fehlt regelmäßig die legitimierende Rechtsgrundlage. Für das Fehlen der Rechtsgrundlage macht es keinen Unterschied, ob die Weitergabe entgeltlich oder unentgeltlich erfolgt. Eine Interessenabwägung trägt nicht, da der Grundsatz zur vertraulichen Behandlung von Personaldaten den berechtigten Interessen des Arbeitgebers überwiegt. Eine Einwilligung der Mitarbeiter in die Weitergabe ist denkbar. Da im Regelfall nicht alle Mitarbeiter zustimmen, werden nur wenige Datensätze übermittelt werden können.

Eine Anonymisierung vor Weitergabe könnte durch eine Interessenabwägung legitimiert werden, da die Anonymisierung die Eingriffstiefe in das Persönlichkeitsrecht der Mitarbeiter deutlich reduziert. Um die Interessenabwägung nutzen zu können, muss der Zweck der anonymen Weitergabe bereits bei der Erhebung der betroffenen Daten festgelegt und den Mitarbeitern im Rahmen der Datenschutzinformation bekannt gegeben worden sein. Andernfalls ist zu prüfen, ob der neue Zweck der Weitergabe kompatibel mit dem ursprünglichen Zweck ist. Das Prüfverfahren beschreibt Art. 6 Abs. 4 DS-GVO. Sofern die Prüfung positiv ausfällt, müssen alle betroffenen Mitarbeiter grundsätzlich über den neuen Zweck informiert werden.

Um Gehaltsstrukturen vergleichen zu können, müssen die relevanten Einflussfaktoren erfasst werden. Dazu zählen u.a. das Alter, die Position, die Qualifikation, das Geschlecht von Mitarbeitern. Bereits die Position ist insbesondere bei Führungskräften personenbezogen. Es gibt im Regelfall nur einen Leiter HR. Doch auch die Kombination aus Alter und Geschlecht kann personenbezogen sein, wenn bspw. unter den 30-40-Jährigen nur eine Frau zu finden ist.

Die Strategie, alle Merkmale zu löschen, die zu einer Identifikation beitragen können, würde faktisch (fast) alle Merkmale löschen. Es bliebe kein nutzbarer Datensatz übrig. Um eine Anonymisierung zu erreichen, bietet es sich daher an, auf andere Anonymisierungsverfahren auszuweichen. Ein Beispiel hierfür wäre die Erzeugung synthetischer Daten auf Basis der statistischen Eigenschaften der dem Synthesemodell zugrundeliegenden personenbezogenen Daten.

8.2.2 Beispiel: Weitergabe von Verkaufszahlen nach Produktkategorien

Eine Versandhändlerin bietet Waren aus unterschiedlichen Produktkategorien an. Sie stellt fest, dass Kunden aus verschiedenen Käufergruppen bereit sind, unterschiedliche Preise für dieselben Produkte zu bezahlen. Sie möchte diese Informationen jetzt erfassen und an Drittparteien vermarkten, um den eigenen Umsatz zu stärken.

Zunächst beginnt sie damit, die Toleranzwerte der jeweiligen Nutzergruppen zu bestimmen. Dazu variiert sie die Preise ihrer verkauften Produkte zufällig und erfasst bei Kaufabschluss einige zusätzliche Parameter dieser und der dazugehörigen Käufer.

Zu diesen Parametern gehören die folgenden personenbezogenen und damit verknüpften Daten: Zu einer Rechnungsadresse und einem erworbenen Produkt werden der Preis, der Kaufzeitpunkt mit Datum und Uhrzeit, die dabei verwendeten Geräte, Browser und Internetanbieter mit IP-Adressen gespeichert.

Als **Angreifer** in diesem Szenario können sowohl interne als auch externe Parteien gesehen werden, die ein Interesse an einer detaillierten Profilerstellung haben könnten. Eine Nutzung könnte dann zu unerwünschter personalisierter Werbung, zum Phishing oder nachfolgendem Identitätsdiebstahl genutzt werden. Auch wenn bei der Erfassung keine Namen von Käufern gespeichert werden, ist es insbesondere einem internen Angreifer möglich, vorhandene Daten (“Hintergrundwissen”) zu nutzen, um Datensätze wieder eindeutig einer Person zuzuordnen. Auch der Käufer eines solchen Datensatzes könnte die Re-Identifikation einer Person oder des gesamten Datensatzes zum Ziel haben. Er könnte damit dann nicht nur seine eigenen Preise - das ursprüngliche Ziel - anpassen, sondern auch die erhaltenen Daten mit seiner eigenen Kaufhistorie vergleichen, um Personen anhand von Kombinationen von erfassten zusätzlichen Merkmalen zu re-identifizieren.

Da bereits die Erfassung dieser Daten zu einem anderen als dem ursprünglichen Zweck der Bestellabwicklung nicht zulässig ist, müssen hier Anonymisierungslösungen angewendet werden, die trotzdem die ursprüngliche Nutzung als Preismechanismus ermöglichen. Daher speichert die Versandhändlerin bei einem Kaufabschluss die Daten nicht

direkt, sondern reduziert sie auf Kategorien unter Entfernung aller direkten personenbezogenen Merkmale. Dabei achtet sie insbesondere darauf, dass für keines der Merkmale eine Übereinstimmung mit weniger als *k* weiteren Datensätzen besteht. Ein Schema für die oben erfassten Daten könnte nun wie in der nachfolgenden Tabelle dargestellt aussehen, wobei die letzte Zeile entsprechend oft für alle Varianten der oberen Parameter wiederholt wird.

Parameter	**Transformation**	**Beispielwert**
Produkt	Ersetzen durch Produktkategorie mit ähnlichem Kaufverhalten	Samsung Galaxy Tab S7 ↪ Oberklasse-Tablet, >2020
Rechnungs-adresse	Ersetzen durch Ortsbereich, der *k* weitere Einträge der Tabelle abdeckt, z.B. mit Open Location Code oder eine allgemeinere Stadt(teil)-angabe	Trankgasse 2, 50667 Köln ↪ 9F28WXR4+ oder: Innenstadt, Köln
Datum, Uhrzeit	Ersetzen durch relevante Kategorien	Abends, Herbst, kein Feiertag, kein Ferientag
Gerät	Ersetzen durch Kategorie	iPhone 12 Pro ↪ Apple-Smartphone
Browser	Streichen von spezifischen Merkmalen	Google Chrome 104 (Windows) ↪ Google Chrome, letzte Version
IP-Adresse	Anonymisierung durch Ersetzen mit dem zugehörigen ISP	84.128.17.3 ↪ Deutsche Telekom
Preis	Angabe der möglichen Preisdifferenz in %	700 € ↪ +5 % des OVP

Tabelle 7: Anonymisierung der personenbezogenen Daten als Vorbereitung der Weitergabe.

8.2.3 Beispiel: Anonymer Abgleich von geleakten Zugangsdaten

Ein Onlinedienst, bei dem sich registrierte Nutzer mit einer E-Mail-Adresse und einem Passwort anmelden, möchte die Sicherheit seiner Plattform verbessern. Dazu möchte er bei der Anmeldung seiner Nutzer die dafür verwendeten Zugangsdaten mit einer Liste von gestohlenen und veröffentlichten Zugangsdaten (genannt Leaks) vergleichen, um seine Nutzer zu warnen und zu verhindern, dass ein unbefugter Angreifer diese Zugangsdaten zur Anmeldung an seinem Dienst nutzen kann.

Listen von solchen im Internet veröffentlichten oder verkauften gestohlenen Zugangsdaten werden von unterschiedlichen Dienstleistern zu einem Abgleich gesammelt. Eine sehr einfache Art und Weise des Abgleichs wäre, dass der Onlinedienst seinem Dienstleister bei einer Anmeldung die von seinen Nutzern verwendeten Zugangsdaten - die E-Mail-Adresse und das unverschlüsselte Passwort - übersendet, damit dieser diese mit seiner Liste vergleichen kann. Dabei hätte der Dienstleister nun direkten Zugriff auf personenbezogene Daten, deren Verarbeitung der ursprüngliche Nutzer nicht zugestimmt hat und die zusätzlich miteinander dienstübergreifend verknüpft und Profile über einen Kunden erstellt werden könnten.

Im Forschungsprojekt EIDI unter der Leitung der Universität Bonn wurde ein Protokoll entwickelt und umgesetzt, bei dem keine personenbezogenen Daten im Klartext an den Dienstleister übertragen werden, aber trotzdem eine zuverlässige Aussage getroffen werden kann, ob sich die Zugangsdaten in einer Leak-Datenbank befinden.

Während eines Login-Versuchs nutzt der Onlinedienst, bei dem sich der Nutzer anmelden möchte, ein kryptographisches Hashverfahren, mit dem die E-Mail-Adresse des Nutzers in eine zufällig aussehende Zeichenfolge umgewandelt wird. Aufgrund der Eigenschaften solcher Funktionen ist es weiterhin möglich, die E-Mail-Adresse einem vollständigen Hash zuzuordnen, was die Anonymität gefährden würde. Wird ein so erzeugter Hash nun abgeschnitten und nur wenige Stellen (Präfix genannt) davon an den Leak-Dienstleister weitergegeben, lässt sich hier k-Anonymität umsetzen. Die Länge dieses Präfixes ist so gewählt, dass sich in der Datenbank immer mindestens k weitere Einträge befinden, die dieses Präfix besitzen, und daher k-Anonymität gewährt wird. Der Leak-Dienstleister kann nun in seiner Datenbank nach Einträgen mit gehashten E-Mail-Adressen suchen, die mit dem erhaltenen Präfix beginnen.

Um nun die Passwörter zu vergleichen, die in der Datenbank des Leak-Dienstleisters gespeichert sind, nutzen beide Seiten ein Verschlüsselungsverfahren basierend auf elliptischen Kurven, um das gewählte Passwort mit zwei Schlüsseln zu verschlüsseln. Zunächst nutzt der Onlinedienst seinen geheimen Schlüssel, um das Passwort zu verschlüsseln und sendet dieses Chiffrat an den Leak-Dienstleister, der dieses ein weiteres Mal mit seinem geheimen Schlüssel verschlüsselt und zurücksendet. Aufgrund der mathematischen Eigenschaften des gewählten Verschlüsselungssystems kann der Onlinedienst jetzt “seine” Verschlüsselung entfernen und erhält das Passwort so, als wäre es nur vom Leak-Dienstleister verschlüsselt worden, ohne dass dieser es aber zur Verschlüsselung lesen konnte.

Der Leak-Dienstleister kann nun die Treffer der Abfrage in der Liste der Leaks mit seinem geheimen Schlüssel verschlüsselt an den Onlinedienst senden, welcher dann die k-anonymisierten Treffer mit dem verschlüsselten Passwort des Nutzers abgleichen kann, ohne dieses zu kennen. Dieser Vergleich ist jetzt möglich, da jeweils derselbe Schlüssel genutzt wurde. Für den Fall eines Treffers kann der Onlinedienst dann weitere Maßnahmen ergreifen, um das Konto des betroffenen Nutzers zu schützen und ihn zu informieren.

8.2.4 Beispiel: Kraftstoffverbrauch von Fahrzeugen

Moderne Autos neuerer Baujahre zeichnen sich durch einen größer werdenden Komfort- und Sicherheitsgewinn aus. So werden beispielsweise Multifunktionslenkräder verbaut, die die Bedienung des Autos angenehmer machen oder Vorgänge, wie das Einschalten von Scheinwerfern und Scheibenwischern automatisiert. Gleichzeitig werden auch mehr Sicherheitssysteme, wie z.B. eine größere Anzahl Airbags verbaut.

Darüber hinaus existieren Sicherheitssysteme wie der eCall[22], welche im Notfall automatisch einen Notruf auslösen und zugleich per Mobilfunk- und GPS- Ortung die Position des Fahrzeugs übermitteln können. Zudem verlangt die EU bei neueren Wagen ein sogenanntes On-Board Fuel Consumption Monitoring (OBFCM) um den Kraftstoffverbrauch zu überwachen und mit den Herstellerangaben zu vergleichen.[23]

Angreifer bei dieser Weitergabe der Daten können vielerlei Art sein. Beispielsweise könnten Versicherungen an einer Identifizierung der Fahrer interessiert sein, da sich hierbei interessante Informationen über das Fahrverhalten des Fahrer generieren lassen.

Betrachtete personenbezogene Daten sind unter anderem das Nutzungsprofil und der Fahrstil, wie Analysen des ADAC zeigen[24][25].
Einzeln aufgegliedert wird das Ausmaß der Datenspeicherung deutlicher. Neben einer Speicherung der gefahrenen Kilometer jeweils getrennt nach Stadt, Land und Autobahn und einzelner Fahrstrecke, werden auch der Kraftstoffverbrauch und die Geschwindigkeit sekundengenau erfasst. Dazu kommt eine Speicherung der GPS-Daten sowie weitere wichtige Fahrzeugdaten.
Der Fahrstil wird anhand der Anzahl der Gurtstraffungen, Motordrehzahl und -temperatur bzw. Daten der Antriebsbatterie sowie Nutzung der verschiedenen Betriebsmodi, z.B. Sportmodus, analysiert.

Eine **Verarbeitung der Daten** kann zum vorgesehenen und von der EU-Verordnung geforderten Zweck erfolgen. Somit macht eine Analyse des Kraftstoff- oder Stromverbrauch nach Stadt, Land und Autobahn Sinn, da diese ebenfalls in der von Herstellern geforderten Verbrauchsangabe geliefert werden müssen.

Die weitere Erfassung des Fahrstils lässt Rückschlüsse auf die Nutzung des Autos zu. Somit ist bei einem sportlicheren Fahrstil ein höherer Verbrauch zu erwarten als bei defensivem Fahren. Dies könnte zur Korrektur oder Einordnung der Verbrauchswerte verwendet werden.

[22] https://www.verbraucherzentrale.de/wissen/reise-mobilitaet/unterwegs-sein/ecall-so-funktioniert-das-automatische-notrufsystem-im-auto-32100 (letzter Zugriff am 28.11.2022).
[23] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/obfcm/ (letzter Zugriff am 28.11.2022).
[24] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/obfcm/ (letzter Zugriff am 28.11.2022).
[25] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/daten-modernes-auto/ (letzter Zugriff am 28.11.2022).

Auch dienen die Daten zum Komfort des Fahrers, indem die Verbrauchswerte auf das Display übermittelt werden und somit nicht notwendigerweise, wie früher üblich, per Dokumentation und Ausrechnen bei Tankstopps ermittelt werden müssen.
Die Funktion der Lokalisierung ist für den eCall notwendig und kann in diesem Fall sogar Leben retten.

Anonymisierungslösungen sind bei Betrachtung der aufgezeichneten Daten offensichtlich unvermeidlich, da bei Auslieferung der Daten in die falschen Hände ein gläserner Fahrer entsteht, dessen Wohn- und Arbeitsort sowie weitere Aufenthaltsorte anhand des Bewegungsprofils einfach bestimmt werden könnten.
Die Fahrleistungen können nach “Stadt”, “Land”, “Autobahn” kategorisiert werden, während der Durchschnittsverbrauch diesen Kategorien zugeordnet wird. Zum Vergleich muss der Fahrzeugtyp mitgeliefert werden. Die Aufzeichnungen einzelner Fahrstrecken werden unterdrückt, ebenso GPS-Daten.

Für weitere Analysen können Werte der aufgezeichneten Parameter festgelegt werden, nach denen Fahrer in “Raser”, “sportlicher Fahrer”, “durchschnittlicher Fahrer”, “defensiver Fahrer” kategorisiert werden können.

Die **Eigenschaften der anonymisierten Daten** können weiterhin die geforderten statistischen Analysen bedienen und den geforderten Abgleich mit den Verbrauchsdaten der Hersteller liefern. Sogar eine Trennung nach Stadt, Land und Autobahn bleibt möglich, ebenso wie eine weitergehende Analyse des Verbrauchs, der das Fahrprofil des Fahrers berücksichtigt.

8.3 Anonymisierung beim Training von Algorithmen

Im Rahmen der Anonymisierung von Trainingsdaten dürfen die Zusammenhänge in den Daten, die die zu erkennenden Muster repräsentieren, nicht verändert werden. Voraussetzung für die Auswahl geeigneter Verfahren zur Anonymisierung ist folglich, dass bekannt ist, welche Zusammenhänge in den Daten für die zu erkennenden Muster wesentlich sind.

Datenschutzrechtlich stellt sich das Training von Algorithmen aus mehreren Perspektiven als problematisch dar. Zu diesem Ergebnis kam bereits eine Studie[26] des Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (Bitkom). Grundsätzlich benötigt das präzise Training eines Modells eine breite Datenbasis, um genügend Informationen bereitzustellen. Andernfalls besteht die Gefahr, dass ein zu grobes Modell erstellt wird. Soll der Algorithmus mit Hilfe des gelernten Modells im späteren Einsatz Autos erkennen und es wurden nur Bilder von SUVs zum Training verwendet, könnte es passieren, dass auch Busse als Autos klassifiziert werden, während Klein- oder Sportwagen außen vor bleiben.

Im Rahmen des Datenschutzes oder des Schutzes von Firmengeheimnissen ist eine Zusammenführung großer Datenmengen jedoch oftmals problematisch, da diese im

[26] Anonymisierung und Pseudonymisierung von Daten für Projekte des maschinellen Lernens - Eine Handreichung für Unternehmen, Bitkom 2020.

Unternehmen, welches das Erstellen von Modellen Künstlicher Intelligenz anbietet, bekannt werden und somit in fremde Hände gelangen. An dieser Stelle bietet sich beispielsweise Federated Learning an.

8.3.1 Beispiel: Federated Learning

Im Grundsatz bedeutet **Federated Learning**, dass die Daten sowohl zum Training genutzt werden können als auch beim jeweiligen Dateneigentümer verbleiben. Der Diensteanbieter erstellt zunächst ein initiales Modell, welches er an seine Partner weitergibt. Unter Verwendung seiner jeweiligen Daten testet nun jeder Partner das erhaltene Modell und teilt dem Diensteanbieter mit, wie die Parameter verändert werden sollen, um das Modell zu verbessern. Aus allen erhaltenen Rückmeldungen errechnet der Diensteanbieter nun ein Gesamtupdate und wendet es auf das bisherige Modell an, welches nun erneut verteilt wird. Direkt erkennbar wird hier die Analogie zur bereits eingeführten rundenbasierten Optimierung des maschinellen Lernens. Vorteilhaft ist hierbei jedoch, dass die Daten beim jeweiligen Eigentümer verbleiben.[27]

Varianten des Federated Learning werden derzeit in der Forschung erprobt. Praxisrelevante Einsätze sind bisher leider nicht bekannt.

Allerdings reicht die Anwendung von Federated Learning aus datenschutzrechtlicher Sicht noch nicht aus. Anhand seiner Antworten auf gestellte Fragen kann ein auf Echtdaten trainierter Algorithmus Informationen über die Trainingsdaten verraten. So können sich Antworten zu im Training verwendeten Datensätzen von Antworten zu neuartigen Datensätzen (minimal) unterscheiden, was zur Aufdeckung personenbezogener Daten führt.

8.3.2 Beispiel: Differential Privacy

Um das in Kapitel 8.3.1 beschriebene Problem der Aufdeckung personenbezogener Daten durch Beobachtung von Veränderungen der Auswertung von Datensätzen zu lösen, kann **Differential Privacy** angewendet werden. Einerseits kann Differential Privacy auf Originaldatensätze angewandt werden, bevor diese zum Training verwendet werden. Hierbei ist zu beachten, dass die daraus erstellten Datensätze weiter für eine gute Modellqualität sorgen. Es entsteht eine „Trade-Off-Situation“. Andererseits ist die Anwendung von Mechanismen des Differential Privacy auf die Ausgaben des Algorithmus möglich.

Auch hierbei handelt es sich um aktuelle Forschungsthemen, deren Eignung für die Praxis sich erst noch zeigen muss.

[27] Anonymisierung und Pseudonymisierung von Daten für Projekte des maschinellen Lernens - Eine Handreichung für Unternehmen, Bitkom 2020.

8.3.3 Synthetische Daten

Ein weiterer Ansatz ist der gänzliche Verzicht auf Realdaten beim Training von Algorithmen. Hierzu können **synthetisierte Daten** verwendet werden. Elementar wichtig ist hier ein gutes Synthesemodell, welches unter Umständen selbst wiederum trainiert werden muss. Hierbei handelt es sich um algorithmisch generierte synthetische Daten. Geht es beispielsweise um die Vorhersage von Krankheiten für gewisse Bevölkerungsgruppen, ist darauf zu achten, dass statistische Gegebenheiten intakt bleiben. Die Gefahr besteht darin, dass in der Realität nicht vorkommende Häufungen gewisser Erkrankungen das Modell unbrauchbar machen. Zudem schützt selbst das Verwenden synthetischer Daten nicht vollends vor der Aufdeckung von Individuen, die einst zur Erstellung des Synthesemodells verwendet worden sind. Auch hierbei muss sich erst die Praxiseignung hinsichtlich Datenschutz und vor allem Modellgenauigkeit zeigen.

8.4 Testen von Software

Die Entwicklung von Software ist mit umfangreichen Tests verbunden. Mit der Durchführung der Tests und der Evaluierung der Testergebnisse soll u.a. sichergestellt werden, dass die gewünschte Funktionalität in der Software korrekt umgesetzt ist. Zum Testen von Software werden sogenannte Testdaten verwendet. Dies sind Daten, die den in der Produktivumgebung zu nutzenden Daten zumindest ähneln. Wird z.B. eine bereits im Einsatz befindliche Software neu entwickelt, so können Echtdaten zum Testen der neu entwickelten Software verwendet werden. Dies ist jedoch mit starken Einschränkungen der DS-GVO verbunden, wenn die Echtdaten personenbezogen sind. In solchen Fällen ist es ratsam, auf synthetische Daten zurückzugreifen. Damit das Testergebnis das zu erwartende Verhalten der Software zur Einsatzzeit hinreichend genau abbildet, müssen die Testdaten den Echtdaten hinsichtlich bestimmter Eigenschaften ausreichend ähneln. Hierzu führt das Entwicklungsteam eine Einschätzung der wichtigsten Eigenschaften der Echtdaten mit erwarteter Auswirkung auf das Verhalten der Software durch. Um die synthetischen Daten möglichst passgenau generieren zu können, werden die als wichtig eingeschätzten Eigenschaften bei der Modellierung der synthetischen Daten nachgebildet.

Bei der Erzeugung von Testdaten unter Berücksichtigung von Eigenschaften echter personenbezogener Daten ist zu beachten, dass die Testdaten die Echtdaten nicht derart nachbilden, dass eine Re-Identifizierung Betroffener durch Nutzung der Testdaten möglich wird. Im Folgenden wird die Erzeugung von Testdaten an Beispielen veranschaulicht.

8.4.1 Softwareaktualisierung/-migration

Das in Unternehmen X seit vielen Jahren eingesetzte Personalverwaltungssystem PVS weist alle von X gewünschten Funktionalitäten auf. Hierzu zählen z.B.

- die Buchhaltung mit Gehaltsabrechnungen
- die Auszahlung der Gehälter
- die Verwaltung der Stammdaten der Mitarbeiter
- die Urlaubsplanung.

Darüber hinaus sind die Mitarbeiter der X gut mit der Nutzung des PVS vertraut. PVS wird von der Firma Y vertrieben und gewartet.

Mittlerweile ist bekannt, dass PVS schwerwiegende Sicherheitslücken aufweist. Firma Y meldet, dass diese Lücken ohne eine grundlegende Überarbeitung der Code-Basis des Systems nicht zu beheben sind. X würde das PVS gerne weiter nutzen und entscheidet daher, Y mit der Überarbeitung zu beauftragen.

Das PVS wird von Y vollständig neu entwickelt. Die ursprünglich im PVS vorhandenen Funktionalitäten werden nachgebildet. Auch die Nutzeroberfläche ist der des alten PVS nachempfunden. Es kommt lediglich moderne, im Vergleich sichere Technologie zum Einsatz.

Nun möchte Y das neu entwickelte PVS in unterschiedlichen Phasen testen. Hierzu sollen u.a. alle Vorgänge nachsimuliert werden, die laut Dokumentation auf dem alten System in den letzten zwei Jahren unter Nutzung der personenbezogenen Daten der Kunden und Mitarbeiter durchgeführt wurden. Im Folgenden sind einige Beispiele gelistet:

- Ein Mitarbeiter A wurde zum 01.10.2021 eingestellt. Seine Lohnsteuerklasse lautet 1, er ist wohnhaft im Bonnweg 123 in Köln und russischer Staatsbürger. Sein Gehalt beträgt EUR 50 000 p.a. Die Stelle ist bis zum 01.12.2024 befristet.
- An alle Mitarbeiter wurde zum 01.12.2021 eine Corona-Einmalzahlung in Höhe von EUR 1000.- brutto überwiesen.
- Mitarbeiter A hat seinen Jahresurlaub 2022 für die Zeit 25.03.-06.05. angemeldet.
- Das durchschnittliche Gehalt eines der 50 Mitarbeiter betrug in 2021 EUR 48.000.- p.a.
- Der Geschäftsführer hatte ein Jahresgehalt von EUR 120.000,-. Er erzielte damit das höchste Jahresgehalt.
- Der durchschnittliche Umsatz, der durch einen der 250 Kunden in 2021 generiert wurde, betrug EUR 20.000,-

Aus Gründen des Datenschutzes kann das Unternehmen X die personenbezogenen Daten der Kunden und Mitarbeiter nicht an das Unternehmen Y weitergeben.
Ein möglicher **Angreifer** könnte durch Zugriff auf die Daten Details aus dem Privatleben der Betroffenen erfahren, die einen schwerwiegenden Eingriff in sein Recht auf informationelle Selbstbestimmung darstellen könnten.

Dieser Angreifer könnte z.B. in der Rolle des Software-Testers auftreten und bei der Validierung der Testdaten unerlaubt Einblick erhalten. Da die Unternehmen X und Y beide im Raum Köln/Bonn angesiedelt sind, ist es zudem nicht unwahrscheinlich, dass sich die Mitarbeiter und Kunden persönlich kennen.

Zu den im PVS **betrachteten personenbezogenen Daten** eines Mitarbeiters zählen Name, Adresse, Lohnsteuerklasse, Nationalität, das Jahresgehalt, das Datum des Arbeitsvertrags und der Zeitraum des Jahresurlaubs.

Weitere Daten sind der durchschnittliche Gewinn pro Kunde, das durchschnittliche Jahresgehalt und der Zeitpunkt und der Betrag der Auszahlung der Corona-Einmalzahlung. Die **geplante Verarbeitung** der Daten umfasst Aktionen, die im Rahmen des Testens des neuen PVS nachgebildet werden sollen. Dies sind

- Das Anlegen eines Datensatzes für einen neuen Mitarbeiter: Hier werden die personenbezogenen Daten erhoben und gespeichert.
- Das Erfassen der Urlaubsplanung eines Mitarbeiters: Hier wird der Zeitraum, die Anzahl der Urlaubstage als zusätzlicher, dem Mitarbeiter zugeordneten Eintrag gespeichert.
- Das monatliche Anstoßen der Überweisung der Gehälter
- Das Anstoßen der Überweisung der Corona-Einmalzahlung
- Die Berechnung der durchschnittlichen Jahresgehälter
- Die Differenz zwischen dem höchsten Jahresgehalt und dem durchschnittlichen Jahresgehalt.
- Die Berechnung des Durchschnitts der Umsätze eines Jahres pro Kunde aus der Summe der in einem Jahr generierten Umsätze.

Im produktionsreifen PVS würden die personenbezogenen Daten aus der entsprechenden Datenbank (hier PVS-DB) extrahiert werden um dann in einzelne Aktionen bzw. Berechnungen einzugehen. Um diese Situation nachzubilden, benötigt das Unternehmen Y also Daten, die den originalen personenbezogenen Daten nachempfunden, jedoch nicht mit jenen identisch sind.

Eine **Anonymisierungslösung** ist in diesem Falle die künstliche Erzeugung von Testdaten, die Eigenschaften der personenbezogenen Daten nachbilden. Hierzu wird eine Datenbank mit denselben Attributen der PVS-DB erzeugt. Die Werte werden unter Nutzung künstlich erzeugter Datensammlungen gesetzt. Im Folgenden werden geeignete Beispiele für die Erzeugung künstlicher Daten anstelle personenbezogener Daten beschrieben.

- Name: Hier können öffentlich verfügbare Listen gängiger Familiennamen verwendet werden.
- Vorname: Hier können Namenslisten von Standesämtern verwendet werden. Dies sind Listen, die Vorschläge für weibliche und männliche Vornamen beinhalten.
- Adresse: Hier kann nach dem folgenden Muster vorgegangen werden, wobei die Erhaltung des Formats der Einträge der PVS-DB zu beachten ist: Vorname, Nachname, Straße, aktuelle Zeilennummer, zufällige fünfstellige Zahl als Postleitzahl, zufällige Auswahl einer Stadt aus dem Gebiet Köln/Bonn
- Lohnsteuerklasse: Zufällige Auswahl von Steuerklassen 1-6
- Jahresgehalt: Zufällige Auswahl eines Wertes, der höher als der aktuelle Mindestlohn ist.
- Eintrittsdatum: Zufälliges Datum
- Geplanter Austritt: Zufälliges Datum.

Eigenschaften der anonymisierten Daten: Das Ergebnis ist eine Datenbank, deren Einträge auf Basis von personenbezogenen Daten anonym erstellte Daten sind, die syntaktisch und strukturell den personenbezogenen Daten gleicht, jedoch keinen Bezug zu einer identifizierbaren natürlichen Person aufweist. Somit wurde verhindert, dass in den Echtdaten vorkommende Kombinationen aus Namen, Geburtsdatum, Staatsangehörigkeit und Adresse „unverfremdet“ in die Testdaten übernommen werden. Dies macht eine Re-Identifizierung Betroffener aus den Testdaten unmöglich.

Hinweis: Die Berechnungen auf den anonymen Daten erfolgen analog zu den auf den Originaldaten ausgeführten Berechnungen.

8.4.2 Funktionalitätstests

Betrachtet man nun das vorangegangene Beispiel, so kann ein Testen der Funktionalität von privilegierten Nutzern des Systems mit bestimmten Berechtigungen ebenfalls erforderlich sein. Die Erzeugung der in Kapitel 8.4.1 beschriebenen Testdaten reicht für diese Funktionalität nicht aus, da die Ebene der Zugriffsberechtigungen untersucht werden soll. Jedoch sind auch die Produktivdaten der Nutzer-Accounts inklusive Login-Credentials und Einträge der Berechtigungen **personenbezogen**. Damit ist eine zweckentfremdete Nutzung auch außerhalb des Produktivsystems regelmäßig nicht zulässig. Insbesondere eine Weitergabe dieser Information an das Unternehmen Y ist nicht zulässig. Ein **Angreifer** könnte die Nutzer-Accounts zum unbefugten Zugriff auf das PVS nutzen und damit Zugriff auf eine Vielzahl sensibler Information erhalten.

Eine **Anonymisierungslösung**, mit der das Testen dieses Teils der Funktionalität der Software möglich wird, beinhaltet das Bereitstellen eines Testsystems, das selbst keinen Zugriff auf echte personenbezogene Daten des PVS hat. Auf diesem Testsystem können nun Nutzeraccounts mit den verschiedenen Berechtigungen, die die echten privilegierten Nutzer haben, angelegt werden. Jedoch werden diese Daten mit synthetischen Daten an den Positionen angereichert, an denen in den Echtdaten personenbezogene, möglicherweise sensible, einer natürlichen Person zuordenbare Information stehen würde. Diese Nutzerdaten können nun im Rahmen von Tests für die Überprüfung der korrekten Funktionalität der Nutzerberechtigungen und deren Einstellungen im PVS verwendet werden.

9. Sonstige rechtliche Anforderungen

9.1 Dokumentationspflichten

Die getroffenen Maßnahmen sowie die relevanten Einflussfaktoren zur Festlegung eines angemessenen Anonymisierungsverfahrens sind zu **dokumentieren**. Dies kann entweder über ein eigenständiges **Anonymisierungskonzept** oder über eine allgemeine Beschreibung im Rahmen der **Darlegung technisch-organisatorischer Maßnahmen** für eine Verarbeitungstätigkeit (s. Kapitel 9.2) erfolgen. Das Verfahren sollte

für Außenstehende **transparent** und **nachvollziehbar** sein. Die Umsetzung dieser Vorgabe in der Dokumentation kann anhand der nachfolgenden Fragen überprüft werden:

- Kann das Verfahren hinsichtlich seiner Wirksamkeit überprüft werden?
- Können die im Verfahren angewendeten Maßnahmen hinsichtlich ihrer Umsetzung verifiziert werden?
- Kann die Einhaltung der angewendeten Maßnahmen evaluiert werden?

Alle mit der Umsetzung betrauten Personen sollten in der Lage sein, den Prozess oder die Maßnahme zu verstehen und entsprechend der definierten Vorgaben umzusetzen.

Es bietet sich an, die Dokumentation anhand des **Vorgehensmodells zur Anonymisierung** auszurichten (s. hierzu Kapitel 9.5).

9.2 Verzeichnis von Verarbeitungstätigkeiten

Jeder Verantwortliche i.S.d. Art. 4 Nr. 7 DS-GVO hat ein **Verzeichnis über seine Verarbeitungstätigkeiten (VVT)** zu führen. Eine Verarbeitungstätigkeit kann als Ablauf verschiedener Verarbeitungsschritte verstanden werden, die mindestens einem übergeordneten Zweck dienen (z.B. Bewerbermanagement, Personalmanagement oder Buchhaltung). Die Verarbeitungsschritte stehen in einer inhaltlich fachlichen Beziehung zum übergeordneten Zweck. Die eingesetzten technischen Hilfsmittel, bspw. Softwareprogramme sind bei der Abgrenzung von Verarbeitungstätigkeiten nicht zu berücksichtigen.

Die Anonymisierung personenbezogener Daten ist grundsätzlich als **Verarbeitung** einzuordnen (s. hierzu Kapitel 4.2). Allerdings verfolgt die Anonymisierung regelmäßig keinen **eigenen Zweck**, sondern ist ein Verarbeitungsschritt, der einer **übergeordneten Verarbeitungstätigkeit** dient (z.B. statistische Auswertung eines Nutzerverhaltens). Daher ist zu empfehlen, die Anonymisierung innerhalb einer Verarbeitungstätigkeit entweder als technisch-organisatorische Maßnahme zu beschreiben oder auf ein eigenes Anonymisierungskonzept zu verweisen. Anonymisierung personenbezogener Daten stellt mit Blick auf das VVT folglich keine eigenständige Verarbeitungstätigkeit dar.

Hinsichtlich der gesetzlichen Forderung nach der Durchführung einer **Datenschutz-Folgenabschätzung** für Verarbeitungen, die voraussichtlich ein **hohes Risiko** für die Rechte und Freiheiten natürlicher Personen zur Folge haben, sollte für die jeweilige Verarbeitungstätigkeit eine Risikoanalyse für Rechte und Freiheiten Betroffener durchgeführt werden. Aus einer Anonymisierung ergeben sich nicht per se ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen. Es müssen weitere Tatbestände hinzukommen, die in der Gesamtschau zu einem hohen Risiko führen.

9.3 Datenschutzinformation

Gemäß Art. 13 DS-GVO muss der Verantwortliche dem Betroffenen bereits bei der Datenerhebung umfassend u.a. über die Zwecke der Datenverarbeitung, die Rechtsgrundlage und Empfänger bzw. Empfängerkategorien **informieren**. Eine Benachrichtigungspflicht hinsichtlich dieser Informationen besteht gemäß Art. 14 DS-GVO auch, wenn der

Verantwortliche die Daten nicht direkt beim Betroffenen erhoben hat, beispielsweise von Dritten oder aus öffentlichen Quellen, wie dem Internet.

Da der Vorgang der Anonymisierung eine Datenverarbeitung darstellt (siehe Kapitel 4.2), muss daher bei der Datenerhebung bzw. bei der Benachrichtigung über eine geplante Anonymisierung informiert werden. Diese Information muss nach dem Wortlaut der Art. 13 und 14 DS-GVO auch den Zweck der Verarbeitung, also der Anonymisierung, beinhalten. Wie nachstehende Beispiele zeigen, können diese eigene Zwecke statistischer Natur sein, z.B. zur Aktivitäts-, Kauf- oder Zahlungsanalyse. Rechtsgrundlage der Zulässigkeit der Anonymisierung für diese Zwecke ist i.d.R. der Erlaubnistatbestand der **Interessenabwägung** gemäß Art. 6 Abs. 1 lit. f) DS-GVO.

Der Verarbeitungsvorgang der Weitegabe bereits anonymisierter Daten an Dritte unterfällt nicht mehr der DS-GVO. Über konkrete **Empfänger oder Empfängerkategorien** ist deshalb nicht mehr zu informieren. Auch eine Datenteilung auf Grundlage der EU-Rechtskate ist damit möglich.

Eine Datenschutzinformation gemäß Art. 13 DS-GVO zum Verarbeitungszweck der Anonymisierung könnte wie folgt lauten: „Neben den geschäftlichen Zwecken anonymisieren wir Ihre Daten für statistische Zwecke und Zwecke der Datenteilung auf Grundlage der EU-Rechtsakte“.

Sofern Daten ohne vorherige Datenschutzinformation über die beabsichtigte Anonymisierung gemäß Art. 13 DS-GVO erhoben worden sind, z.B. bei **älteren Datenbeständen**, erlaubt Art. 6 Abs. 4 lit. e) DS-GVO eine **Weiterverarbeitung** bei Vorhandensein geeigneter Garantien. Dazu können nach dem Wortlaut der DS-GVO auch die Verschlüsselung oder die Pseudonymisierung. gehören. Diese Norm ist geschaffen worden, um nach DS-GVO Big **Data-Analysen**, die ergebnisoffen mit dem Ziel der Mustererkennung oder der Generierung neuer (personenbezogener) Informationen und insoweit ohne konkrete oder mit sich verändernder Zweckbestimmung erfolgen, zu ermöglichen. Sofern bereits eine Pseudonymisierung eine geeignete Garantie für eine zweckändernde Nutzung ausreichen kann, ist die Anonymisierung eine weitaus effektivere Garantie. Damit wäre über Art. 6 Abs. 4 auch eine Nutzung älterer Datenbestände nach einer Anonymisierung für Analysezwecke oder zur Datenteilung möglich.

Ob über die nachträgliche Anonymisierung noch informiert werden muss, beurteilt sich gemäß Art. 14 Abs. 5 lit. b) DS-GVO. Diese Regelung stellt auf einen **unverhältnismäßigen Aufwand** ab. Als Beispiel hierfür werden statistische Zwecke genannt. Mit Blick auf das Persönlichkeitsrecht und den Datenschutz entstehen dem Betroffenen nach einer Anonymisierung keine Gefahren mehr. Eine nachträgliche Information über eine Anonymisierung schafft für den Betroffenen keine datenschutzrechtlichen Mehrwerte, sondern erzeugt nur einen unverhältnismäßigen Aufwand.

9.4 Prüfpflichten

Wer Daten anonymisiert hat oder anonymisierte Daten nutzt, ist verpflichtet, **kontinuierlich zu prüfen**, dass die Anonymisierung gewahrt bleibt.[28] Dazu muss er prüfen, ob der Personenbezug wiederhergestellt werden kann. Die Durchführung und das Prüfergebnis sollten dokumentiert werden.

Geprüft wird:

- Ist ein Herausgreifen („singling out“) möglich?
- Ist eine Verknüpfung des Datensatzes der gleichen Person mit Datensätzen aus legal beschaffbaren Datensätzen zu dieser Person statistisch möglich?
- Lassen sich mit einer signifikanten Wahrscheinlichkeit Werte eines Merkmals aus Werten anderer Merkmale im Datenbestand ableiten?

Bei der Prüfung, ob Daten anonym sind, kommt es nicht auf sichere oder richtige Erkenntnisse an. Vielmehr reicht es aus, dass ein Personenbezug mit einer **gewissen Wahrscheinlichkeit** oder für einen Teil der Datensätze wiederhergestellt werden kann. Kommt ein Verantwortlicher bei der Prüfung zu dem Ergebnis, dass eine Person identifiziert oder identifizierbar ist, leben die Anforderungen der DS-GVO an die Datenverarbeitung auf, so auch zur Frage der Rechtmäßigkeit der Verarbeitung. Ist keine Rechtsgrundlage auf Seiten des Verantwortlichen gegeben, sind die personenbezogenen Daten zu löschen. Darüber hinaus sollte geprüft werden, ob durch die Aufhebung der Anonymisierung eine **Meldepflicht** bei der zuständigen Aufsichtsbehörde entsteht bzw. Betroffene informiert werden müssen (vgl. Art. 33 u. 34 DS-GVO).

> **Hinweis:** Auch die Überprüfung der Speicherdauer anonymisierter Daten kann im Einzelfall sinnvoll sein. Immerhin kann ein gelöschter anonymisierter Datensatz nicht mehr Grundlage für eine mögliche Re-Identifizierung Betroffener sein.

9.5 Vorgehensmodell zur Anonymisierung

Um personenbezogene Daten zu anonymisieren, bietet sich folgender Ablauf an:

Lfd. Nummer	Maßnahme	Kapitel im Leitfaden
1	**Ermittlung der Rechtsgrundlage für die Anonymisierung (z.B. in der DS-GVO, im BDSG oder in einem bereichsspezifischen Gesetz)**	4.2
2	**Sicherstellen, dass die Informationspflichten gemäß Art. 13 und 14 DS-GVO umgesetzt werden**	9.3

[28] Stellungnahme 5/2014 der Artikel 29-Gruppe zu Anonymisierungstechniken, WP 216, S. 4.

3	**Auswahl und Festlegung des geeigneten Anonymisierungsverfahrens**	
3.1	Art und Risikoklasse der zu anonymisierenden personenbezogenen Daten	6.1 und 8
3.2	Beabsichtigte Verarbeitungszwecke	6.1 und 8
3.3	Kontext der Anonymisierung	6.1 und 8
3.4	Erwartete Anzahl der Datensätze	6.1 und 8
3.5	Ermittlung der statistischen Eigenschaften in den Datensätzen, die benötigt werden und welche Merkmale für diese Eigenschaften relevant sind	6.3 und 8
3.6	Festlegung des geeignetes Anonymisierungsverfahrens und dessen Zeitpunkt	9.1, 9.2 und 9.4
4	**Durchführung der Anonymisierung**	
4.1	Entfernung aller direkten Identifikationsmerkmale (z.B. Name, Anschrift, Kontaktdaten, Kreditkartennummer)	3.2 und 6.1
4.2	Entfernen aller nicht benötigten indirekten Identifikationsmerkmale (z.B. Geschlecht, körperliche Erscheinungsmerkmale, Alter, Postleitzahl)	3.2 und 6.1
4.3	Durchführung eines oder mehrerer Verfahren der • Randomisierung • Generalisierung oder • Durchführung eines Verfahrens mit synthetischen Daten.	6.36.3.1
5	**Analyse, ob und welche Risiken zum Wiederherstellen des Personenbezugs bestehen**	6.1 und 6.2
6	**Sofern Risiken bestehen, Anwendung weiterer Verfahren zur Anonymisierung**	6.3.1
7	**Schritte 4.1 bis 4.3 durchlaufen, bis keine Risiken mehr erkennbar sind**	6.3.1
8	**Prüfen, ob die benötigten statistischen Eigenschaften erhalten geblieben sind**	6.3
9	**Prüfung und Ergebnis dokumentieren**	9.1
10	**Anonymisierte Daten nutzen oder weitergeben**	
11	**Regelmäßig die Prüfung gemäß Schritt 5 wiederholen, ggf. Schritte 6 und 7 anwenden und gemäß Schritt 9 dokumentieren**	

STIFTUNG DATENSCHUTZ

Stiftung Datenschutz
rechtsfähige Stiftung bürgerlichen Rechts
Karl-Rothe-Straße 10–14
04105 Leipzig
Deutschland

Telefon 0341 / 5861 555-0
mail@stiftungdatenschutz.org
www.stiftungdatenschutz.org

gestiftet von der Bundesrepublik Deutschland
vertreten durch den Vorstand Frederick Richter

Die Arbeit der Stiftung Datenschutz wird aus dem Bundeshaushalt gefördert (Einzelplan des BMJ).

Bundesministerium der Justiz