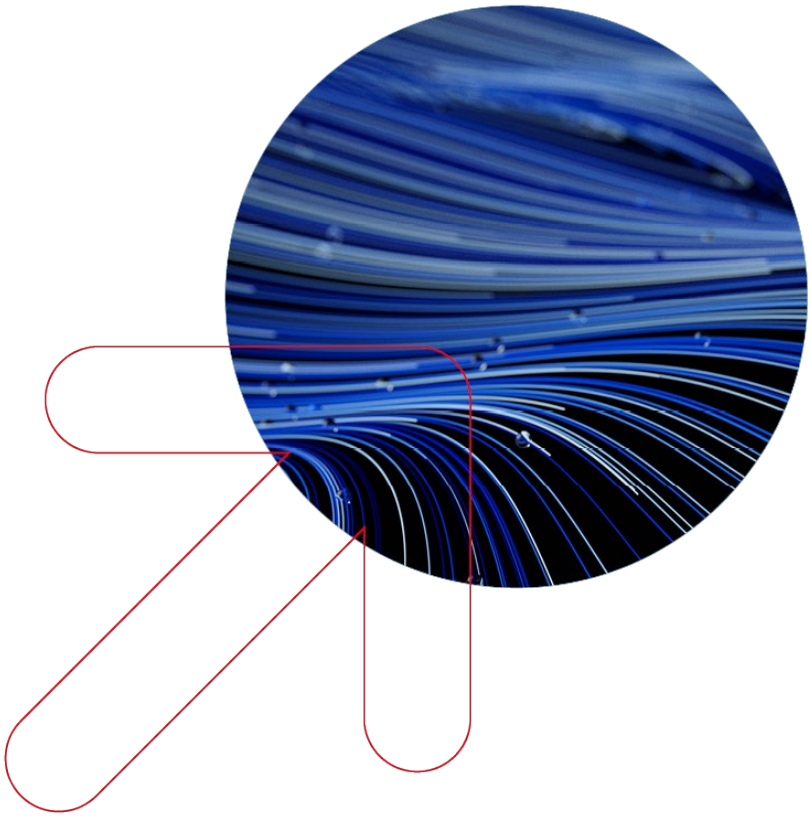
WIK • Diskussionsbeitrag

Nr. 527



Data Access, Data Sharing und Privacy

Autoren:
Andrea Liebe
Peter Kroon
Lukas Wiewiorra

Bad Honnef, Dezember 2024

WIK
Wissenschaftliches Institut
für Infrastruktur und
Kommunikationsdienste

Impressum

WIK Wissenschaftliches Institut für
Infrastruktur und Kommunikationsdienste GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Deutschland
Tel.: +49 2224 9225-0
Fax: +49 2224 9225-63
E-Mail: info@wik.org
www.wik.org

Vertretungs- und zeichnungsberechtigte Personen

Geschäftsführerin und Direktorin	Dr. Cara Schwarz-Schilling
Direktor, Verwaltungs- und Abteilungsleiter	Alex Kalevi Dieke
Direktor, Abteilungsleiter	Prof. Dr. Bernd Sörries
Abteilungsleiter	Dr. Christian Wernick
Abteilungsleiter	Dr. Lukas Wiewiorra
Vorsitzender des Aufsichtsrates	Dr. Thomas Solbach
Handelsregister	Amtsgericht Siegburg, HRB 7225
Steuer-Nr.	222/5751/0722
Umsatzsteueridentifikations-Nr.	DE 123 383 795

Stand: Januar 2024

ISSN 1865-8997

Weitere Diskussionsbeiträge finden Sie hier:
https://www.wik.org/veroeffentlichungen/diskussionsbeitraege

In den vom WIK herausgegebenen Diskussionsbeiträgen erscheinen in loser Folge Aufsätze und Vorträge von Mitarbeitern des Instituts sowie ausgewählte Zwischen- und Abschlussberichte von durchgeführten Forschungsprojekten. Mit der Herausgabe dieser Reihe bezweckt das WIK, über seine Tätigkeit zu informieren, Diskussionsanstöße zu geben, aber auch Anregungen von außen zu empfangen. Kritik und Kommentare sind deshalb jederzeit willkommen. Die in den verschiedenen Beiträgen zum Ausdruck kommenden Ansichten geben ausschließlich die Meinung der jeweiligen Autoren wieder.

Inhaltsverzeichnis

Inhalt

Zusammenfassung

Die Studie beleuchtet das Spannungsfeld zwischen Datenschutz, Nutzbarkeit anonymisierter Daten und Verhältnismäßigkeit verschiedener Anonymisierungsverfahren. Es wird deutlich, dass die Interessen von Dateninhabern und Datenempfängern oft gegensätzlich sind: Während Dateninhaber einen hohen Anonymisierungsgrad bevorzugen, um den Datenschutz zu maximieren, wünschen Datenempfänger eine geringere Anonymisierung, um eine bessere Datenqualität und Nutzbarkeit zu erhalten.

Der Schwerpunkt liegt auf der Analyse verschiedener Anonymisierungsverfahren. Ihre Stärken, Schwächen und die praktische Umsetzung werden gegenübergestellt, wobei die Wahl des geeigneten Verfahrens als stark abhängig von den Anwendungszielen, den Dateneigenschaften sowie den ökonomischen und technischen Rahmenbedingungen beschrieben wird. Einfache Verfahren sind leicht umsetzbar, beeinträchtigen aber häufig die Datenqualität, während moderne Ansätze bessere Kompromisse bieten, aber anspruchsvoller und kostenintensiver sind.

Die Studie unterstreicht, dass ein sinnvoller Datenzugang nur durch die Kombination von geeigneten Anonymisierungsverfahren, klaren regulatorischen Vorgaben und der Berücksichtigung der Interessen aller Akteure erreicht werden kann.

Summary

The study draws attention to the inherent tensions between the protection of data, the usability of anonymised data and the proportionality of the various anonymisation methods employed. It becomes evident that the interests of data owners and data recipients are frequently in conflict. While data owners tend to favour a high degree of anonymisation to maximise data protection, data recipients typically prefer a lower degree of anonymisation to obtain better data quality and usability.

The focus is on analysing different anonymisation methods. The strengths and weaknesses of each method are evaluated, and the choice of an appropriate method is shown to be highly dependent on the specific objectives of the application, the characteristics of the data, and the economic and technical context in which it is to be used. The implementation of simple methods is straightforward; however, they often result in a reduction of data quality. In contrast, modern approaches offer superior compromises but are more demanding and cost intensive.

The study emphasises that meaningful data access can only be achieved through a combination of suitable anonymisation procedures, clear regulatory requirements and consideration of the interests of all stakeholders.

1 Einleitung

Der Austausch von Daten bietet Unternehmen erhebliches Potenzial, um Produktionsprozesse zu optimieren, die Zusammenarbeit in Wertschöpfungsketten zu verbessern und die Produktentwicklung zu beschleunigen. Insbesondere der Zugriff auf umfangreiche Datenbestände, wie sie beispielsweise großen Online-Plattformen zur Verfügung stehen, kann signifikante Wertschöpfung ermöglichen. Vor diesem Hintergrund sehen Regulierungen wie der Digital Markets Act (DMA)[1] es vor, dass Daten mit anderen Akteuren verpflichtend geteilt werden müssen, um Wettbewerbsbedingungen zu verbessern.

Gleichzeitig müssen aber Mechanismen etabliert und regulatorisch überprüft werden, die gewährleisten, dass der ökonomische Wert dieser Daten gehoben wird, ohne die Datenschutzinteressen der betroffenen Personen zu verletzen. Die Anonymisierung von Daten ist ein solcher Mechanismus. Sie hat zum Ziel, personenbezogene Daten so zu verändern, dass sie nicht mehr auf eine bestimmte Person zurückgeführt werden können. Dabei gibt es unterschiedliche Ansätze und Techniken, die je nach Anwendungsfall genutzt werden können.

Das Ziel dieser Untersuchung ist es, das Spannungsfeld zwischen Datenschutz, Verwendbarkeit von Daten und der Verhältnismäßigkeit der Anonymisierung systematisch aufzuarbeiten. Folgende Forschungsfragen sollen dabei adressiert werden:

- Welche ökonomischen Anreize haben die verschiedenen Akteure und wie beeinflussen diese Anreize den Zugang zu Daten (Data Access) und den Datenaustausch (Data Sharing)?
- Welche Anonymisierungsverfahren existieren aktuell? Wo liegen ihre Stärken und Schwächen und welche Verfahren kommen in der Praxis tatsächlich zum Einsatz?
- Welche Verfahren zur Anonymisierung von Daten sind für Gatekeeper im Rahmen des DMA besonders praktisch geeignet um sowohl wirtschaftliche als auch regulatorische Anforderungen zur erfüllen?

Die Untersuchung wird sich an diesen Leitfragen orientieren und gliedert sich entsprechend in die folgenden Abschnitte.

Nach dieser Einleitung wird in Kapitel 2 zunächst der Hintergrund der Untersuchung dargelegt, indem die Notwendigkeit der Anonymisierung von im Kontext von Data Access und Data Sharing erläutert wird. Dies umfasst neben der ökonomischen Perspektive und dem regulatorischen Hintergrund auch die Akteure und ihre Interessenlagen sowie das Risiko einer De-Anonymisierung. Kapitel 3 widmet sich der Struktur und Typologisierung von Daten und geht dabei auf ihre Nutzbarkeit ein. In Kapitel 4 werden verschiedene Verfahren zur Anonymisierung von Daten systematisch vorgestellt und hinsichtlich ihrer Funktionsweise und Anwendbarkeit beschrieben. Dazu gehören grundlegende Ansätze

[1] Europäisches Parlament (2022b).

wie Suppression, Datenmaskierung und Aggregation aber auch etwas komplexere Verfahren wie z.B. Noise Addition, Randomisierung und, Permutation. Ein Schwerpunkt liegt auf den Verfahren der Generalisierung, insbesondere K-Anonymität, L-Diversität und T-Closeness. Ergänzt wird die Betrachtung durch moderne Ansätze wie Differential Privacy, Synthetisierung und Model-based Obfuscation Knowledge (MOK). In Kapitel 5 erfolgen verschiedene Analysen zur Auswahl geeigneter Verfahren zur Anonymisierung von Daten. Dazu werden zunächst Auswahlkriterien definiert, bevor verschiedene vergleichende Analysen der Anonymisierungsverfahren durchgeführt werden. Dazu zählen unter anderem die Beurteilung anhand von Kriterien, die Beurteilung der Eignung Datenart sowie nach Nutzungsszenario. Die Studie schließt mit einer Schlussbetrachtung und einem Ausblick.

2 Notwendigkeit der Anonymisierung von Daten im Kontext von Data Access und Data Sharing

2.1 Ökonomische Perspektive

Daten sind zentraler Bestandteil nahezu aller wirtschaftlichen Aktivitäten und können aus ökonomischer Perspektive als immaterielles Gut betrachtet werden. Sie zeichnen sich durch die Eigenschaft aus, mehrfach und zu unterschiedlichen Zwecken genutzt zu werden, ohne dabei verbraucht zu werden.[2] Ein besonderes Merkmal von Daten ist ihre Nicht-Rivalität im Konsum: Mehrere Akteure können dieselben Daten simultan verwenden, ohne dass dadurch Konkurrenz oder Ausschluss entsteht. Dies bedeutet, dass die Nutzung von Daten durch ein Unternehmen nicht automatisch die Verfügbarkeit oder Nutzbarkeit der gleichen Daten für ein anderes Unternehmen einschränkt.[3]

Im Zuge der Digitalisierung und der Entwicklung aller möglichen internetbasierten Dienste entstehen umfangreiche Datenbestände. Der ökonomische Wert von Daten entsteht durch ihre Verarbeitung und Kombination. Dieser Wertschöpfungsprozess basiert auf Erkenntnissen, die durch Analyse generiert werden. Dabei spielen Dienstleister und Infrastrukturanbieter eine Schlüsselrolle, da sie Datenerfassung, -übertragung, -speicherung, -analyse und -nutzung ermöglichen. Diese datenbasierten Aktivitäten sind in nahezu allen Branchen und Unternehmen zu finden, wenn auch in unterschiedlicher Intensität.[4]

Die Datenanalyse eröffnet Unternehmen die Optimierung ihrer Geschäftsmodelle. Sie kann unter anderem zur (1) Neukundengewinnung, (2) adressatenbezogenen Ansprache, z.B. bei Rabattaktionen, (3) Erhöhung der Kundenbindung durch die Anpassung von Produkten und (4) Steigerung der internen Effizienz genutzt werden. All diese Punkte münden letztlich in Umsatzsteigerungen oder Kosteneinsparungen. Vor diesem Hintergrund hat der Zugang zu diesen Daten also eine zentrale Bedeutung.[5]

Aus wettbewerblicher Perspektive ist der Zugang zu Daten ein zentraler Faktor, der die Marktstruktur, Innovationsdynamik und Wettbewerbsfähigkeit von Unternehmen maßgeblich beeinflusst. Unternehmen mit Zugang zu umfangreichen, hochwertigen Datenbeständen können präzisere Analysen durchführen, ihre Geschäftsprozesse optimieren und personalisierte Dienstleistungen anbieten. Skaleneffekte, die durch effizientere Datenverarbeitung entstehen, und Netzwerkeffekte, die den Nutzen von datenbasierten Dienstleistungen bei wachsender Nutzerzahl erhöhen, verstärken diese Vorteile. Dadurch entstehen erhöhte Markteintrittsbarrieren für neue Akteure, was etablierten Unternehmen einen signifikanten Wettbewerbsvorteil verschafft.[6]

---

2 Vgl. Arnold, R. et al. (2020).
3 Vgl. Hjørland, B. (2018).
4 Vgl. BVDW (2018).
5 Vgl. Arnaut, C. et al. (2018).
6 Vgl. Grunes, A. P./ Stucke, M. E. (2015).

Exklusiver Zugang zu Daten ermöglicht es einem Unternehmen, seine Marktstellung zu festigen und auszubauen. Diese Datenkonzentration birgt die Gefahr einer Monopolisierung, da der Zugang zu Daten eine zentrale Ressource für wettbewerbsfähige Geschäftsmodelle darstellt. Erschwert ein etabliertes Unternehmen anderen Akteuren den Zugang zu Daten, kann es seine Wettbewerbsposition stärken und seine Marktmacht weiter ausbauen.[7] Eine Verpflichtung, Daten zu teilen und Datenzugang zu gewähren, kann dem entgegenwirken, indem es den Wettbewerb durch die Stärkung kleinerer Akteure fördert.[8]

In der Regel erfolgt der Datenaustausch zwischen Unternehmen horizontal auf Grundlage bilateraler vertraglicher Vereinbarungen und gemeinsamer Initiativen von Unternehmen. Gleichwohl werden Daten lediglich innerhalb des eigenen Sektors und auch nur in einem geringen Umfang geteilt. Gründe dafür liegen sowohl in technischen als auch in rechtlichen Hindernissen sowie der Verweigerung des Zugangs aus unterschiedlichen strategischen und wettbewerblichen Gründen.[9]

Im Folgenden wird der regulatorische Rahmen vorgestellt, bevor auf die Herausforderungen, Hindernisse und Spannungsfelder eingegangen wird, die den umfassenden Datenaustausch behindern.

2.2 Regulatorischer Hintergrund

Data Access und Data Sharing sind Teil der Europäischen Datenstrategie. Diese basiert auf der Erkenntnis, dass Daten eine wesentliche Ressource für Wirtschaftswachstum, Wettbewerbsfähigkeit, Innovation, Schaffung von Arbeitsplätzen und gesellschaftlichen Fortschritt im Allgemeinen sind bzw. sein können. Die Datenstrategie hat das Ziel, einen Binnenmarkt für Daten zu schaffen, der die Wettbewerbsfähigkeit und die Datensouveränität in Europa gewährleistet.[10] Sie wird von zahlreichen Regelungen flankiert, auf deren Kern im Folgenden skizzenhaft eingegangen wird.

In Europa bestehen bereits seit geraumer Zeit sektorspezifische Regelungen, welche darauf abzielen, Data Sharing und Data Access zu intensivieren. Als Beispiele können Regulierungen wie die Richtlinie über Zahlungsdienste im Binnenmarkt von 2015[11] oder die Richtlinie zum Rahmen für die Einführung intelligenter Verkehrssysteme im Straßenverkehr von 2010[12] angeführt werden. Darüber hinaus existieren horizontale Regelungen, welche alle Sektoren betreffen und aus verschiedenen Gesetzen und Verordnungen auf EU-Ebene sowie auf nationalen Gesetzgebungen der Mitgliedstaaten beruhen.

---

7 Vgl. Shapiro, C./ Varian, H.R. (2013).
8 Vgl. Graef, I. et al. (2019).
9 Vgl. Batura O. et al. (2023).
10 Vgl. Europäische Kommission (2020).
11 Vgl. Europäisches Parlament (2015).
12 Vgl. Europäisches Parlament (2010).

Zunächst einmal ist die zentrale **Datenschutz-Grundverordnung (DSGVO)**[13] aus dem Jahr 2016, die eine verpflichtende Regelung darstellt, zu nennen. Die DSGVO regelt den Umgang mit personenbezogenen Daten und definiert klare Kriterien für die Rechtmäßigkeit, Transparenz und Fairness der Verarbeitung dieser Daten. Die Freigabe von und der Zugang zu Daten haben stets im Einklang mit den Bestimmungen der DSGVO zu erfolgen, insbesondere im Hinblick auf die Einwilligung, die Zweckbindung und die Rechte der betroffenen Personen.

Des Weiteren existieren horizontale EU-Regelungen, die das Ziel haben, Rahmenbedingungen zu schaffen, die das Teilen von Daten auf einer freiwilligen Basis innerhalb und über Sektorengrenzen hinaus zu stimulieren. Gleichwohl haben diese Aktivitäten unter den Maximen des Schutzes personenbezogener Daten zu erfolgen. Folgende Regulierungen sind in diesem Kontext zu nennen:

Seit dem Jahr 2018 existiert die **Verordnung über die Freizügigkeit von Daten in der EU** (Free Flow of Data)[14], die darauf abzielt, die Nutzung und den Austausch von nicht-personenbezogenen Daten innerhalb der EU zu fördern. Sie soll sicherstellen, dass nicht personenbezogene Daten überall in der Union gespeichert, verarbeitet und übermittelt werden können.

Die **Open Data Direktive**[15] aus dem Jahr 2019 ist ein wesentlicher Bestandteil der europäischen Strategie für Daten. Als Rechtsrahmen für Daten des öffentlichen Sektors hat sie Transparenz und fairen Wettbewerb als zentrale Leitprinzipien. Die Umsetzung dieser Richtlinie auf nationaler Ebene ist für die kommenden Jahre vorgesehen.

Der **Data Governance Act** (DGA)[16] von 2022 ist ein sektorübergreifendes Instrument, das darauf abzielt, einheitliche Rahmenbedingungen für den Datenaustausch innerhalb des Europäischen Binnenmarkts zu schaffen. Sowohl die Verfügbarkeit als auch die Weiterverwendung von Daten soll gefördert werden, indem neue Regeln für klar definierte Bedingungen der Datenteilung, Datentreuhänder und datenaltruistische Organisationen eingeführt werden. Sowohl personenbezogene als auch nicht-personenbezogene Daten fallen in den Anwendungsbereich des DGA.

Neben diesen eher freiwilligen und stimulieren wirkenden Regelungen haben die nachstehenden horizontalen EU-Regelungen einen verpflichtenden Charakter. Zunächst ist der **Data Act**[17] zu nennen, der im Januar 2024 in Kraft getreten ist und Grundregeln für alle Sektoren in Bezug auf Rechte zur Datennutzung aufstellt. Er schafft Prozesse und Strukturen, um die gemeinsame Nutzung von Daten durch Unternehmen, Einzelpersonen und den öffentlichen Sektor zu erleichtern. Ziel ist es, die Datenwirtschaft in der EU

13 Vgl. Europäisches Parlament (2016).
14 Vgl. Europäisches Parlament (2018).
15 Vgl. Europäisches Parlament (2019).
16 Vgl. Europäisches Parlament (2022a).
17 Vgl. Europäisches Parlament (2023).

anzukurbeln, indem Industriedaten geöffnet, ihre Zugänglichkeit und Nutzung optimiert und ein wettbewerbsfähiger und zuverlässiger europäischer Datenmarkt gefördert wird.

Das im November 2022 in Kraft getretene **Gesetz über digitale Märkte** (Digital Markets Act, DMA)[18] hat insbesondere große Onlineplattformen, die auf digitalen Märkten als so gennannte Torwächter fungieren, im Fokus. Es definiert die folgenden Regeln in Hinblick auf Data Sharing und Data Access:

- Gemäß Artikel 5 (2) ist die Nutzung personenbezogener Daten zu kommerziellen Zwecken sowie die Zusammenführung personenbezogener Daten aus dem jeweiligen Kerndienst der Plattform mit Daten aus anderen Diensten des Gatekeepers untersagt, sofern der Endnutzer nicht seine Einwilligung erteilt hat.
- Gemäß Artikel 6 (2) ist die Verwendung nicht öffentlicher Daten von gewerblichen Nutzern (oder deren Kunden), die aus der Nutzung der Kerndienste der Gatekeeper-Plattform (und der kombinierten Dienste) stammen, durch Dritte untersagt, sofern dies geschieht, um mit diesen gewerblichen Nutzern in Wettbewerb zu treten.
- Gemäß Artikel 5 (9,10) besteht die Verpflichtung zur Weitergabe von Daten an Werbetreibende und Anbieter von Online-Werbung oder von diesen beauftragte Dritte. Dies umfasst auch Informationen, auf deren Grundlage die Werbekosten und die Vergütung der genannten Anbieter berechnet werden.
- Gemäß Artikel 6 (8) sind die Torwächter dazu verpflichtet, den Werbetreibenden und Anbietern von Online-Werbung (bzw. von diesen beauftragten Dritten) Zugang zu den für die Überprüfung des Werbeinventars erforderlichen Daten zu gewähren. Dadurch wird diesen ermöglicht, eigene Performance-Tools zur Messung der genutzten Kerndienste der Plattform einzusetzen.
- Gemäß Artikel 6 (9) besteht die Verpflichtung für Torwächter, Endkunden (oder autorisierten Dritten), Echtzeit-Zugriff und Portabilität ihrer Daten bereitzustellen.
- Gemäß Artikel 6 (10) besteht für Torwächter die Verpflichtung, gewerblichen Nutzern einen effektiven, hochwertigen und permanenten Echtzeitzugang zu aggregierten und nichtaggregierten Daten, einschließlich personenbezogener Daten, die im Zusammenhang mit der Nutzung der betreffenden zentralen Plattformdienste bereitgestellt werden, zu gewährleisten. Sofern personenbezogene Daten betroffen sind, ist deren Weitergabe an Endkunden nur mit deren Einwilligung zulässig.
- Gemäß Artikel 6 (11) besteht die Verpflichtung, Drittunternehmen, die Online-Suchmaschinen bereitstellen, Zugang zu Ranking-, Anfrage-, Klick- und Ansichtsdaten zu gewähren, wobei eine Anonymisierung erforderlich ist.

Der DMA stärkt den Wettbewerb und die Transparenz, indem es Gatekeeper als Dateninhaber verpflichtet, Datenzugang und Datenweitergabe unter klar definierten Bedingungen zu ermöglichen. Gleichzeitig schützt es die Rechte von Endnutzern und gewerblichen

18 Europäisches Parlament (2022b).

Nutzern, wobei zweiter in der Regel als Datennachfrager gelten, durch Einschränkungen in der Datennutzung und Anforderungen an die Einwilligung.

Das **Gesetz über digitale Dienste** (Digital Services Act, DSA)[19] bringt einheitliche Regelungen für digitale Dienste innerhalb der EU, mit einem besonderen Fokus auf Transparenz, Rechenschaftspflichten und den Schutz personenbezogener Daten. Im Hinblick auf Data Sharing und Data Access gilt, dass die Übermittlung personenbezogener Daten ohne die Einwilligung der betroffenen Person ausdrücklich verboten ist. Dies umfasst auch Daten, die durch manipulative Inhalte erlangt wurden (Art. 28). Große Online-Plattformen und Suchmaschinen, Very Large Online Platforms (VLOPs) und Very Large Online Search Engines (VLOSEs), sind verpflichtet, das Teilen nicht autorisierter personenbezogener Daten auf ihren Plattformen zu verhindern (Art. 28). Artikel 40 sieht vor, dass VLOPs und VLOSEs Datenzugang für Digital Service Koordinatoren und Forschende gewähren. Dieser Zugang dient der Bewertung der Einhaltung der Verordnung sowie der Identifizierung systemischer Risiken in der EU.

Weitere Regelungen werden perspektivisch den Zugang und das Teilen von Daten bestimmen. Der **AI Act**[20], im August 2024 in Kraft getreten und nun sukzessive anzuwenden, ist der erste Rechtsrahmen speziell für Künstliche Intelligenz. Er hat klare Anforderungen und Verpflichtungen für Entwickler und Betreiber von KI. Ein Kernpunkt in Hinblick auf den Zugang zu Daten ist, dass die Verordnung vorsieht, dass so genannte Hochrisiko-KI-Systeme mit hochwertigen Datensätzen trainiert werden müssen, was wiederum einen angemessenen Datenzugang bei gleichzeitiger Einhaltung der Vorschriften der DSGVO[21] erfordert. Darüber hinaus haben die Anbieter der Hochrisiko-KI-Systeme Transparenzvorschriften einzuhalten. In Artikel 54 werden darüber so genannte regulatorische Reallabore (Sandboxes) vorgesehen, die dem Austausch von Daten in einem geschützten Rahmen dienen sollen.

Darüber hinaus werden technische **Leitlinien für den europäischen Rahmen für digitale Identitäten** erwartet, mit dem Ziel, personenbezogene Daten durch elektronische Identifizierung und Authentifizierung zu schützen. Dies soll durch 'digitale Brieftaschen' (eID-Systeme) geschehen, die die Mitgliedstaaten auf der Grundlage vorgeschlagener gemeinsamer technischer Standards einführen müssen.

2.3 Akteure und ihre Interessenlagen

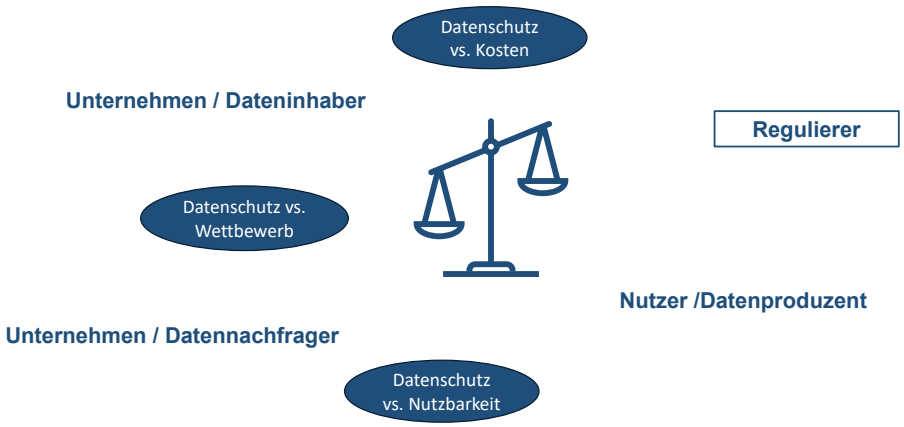
Data Access erfordert in der Regel eine Anonymisierung der Daten. Dabei entsteht ein Spannungsfeld der Interessen der beteiligten Akteure, das hier skizziert werden soll.

---

19 Europäisches Parlament (2022c).
20 Europäisches Parlament (2024).
21 Europäisches Parlament (2016).

Das Spannungsfeld erstreckt sich zwischen den bestehenden Datenschutzanforderungen, den wirtschaftlichen Interessen und der Nutzbarkeit der Daten. Beteiligte Akteure sind zunächst einmal Unternehmen als Dateninhaber auf der einen Seite, Unternehmen als Datennachfrager auf der anderen Seite und die Nutzer von Onlinediensten als Datenproduzenten. Regulierungsbehörden kommen als ausgleichende und abwägende Instanz ins Spiel.

Abbildung 2-1: Akteure und Interessenlagen



Quelle: WIK, Eigene Darstellung.

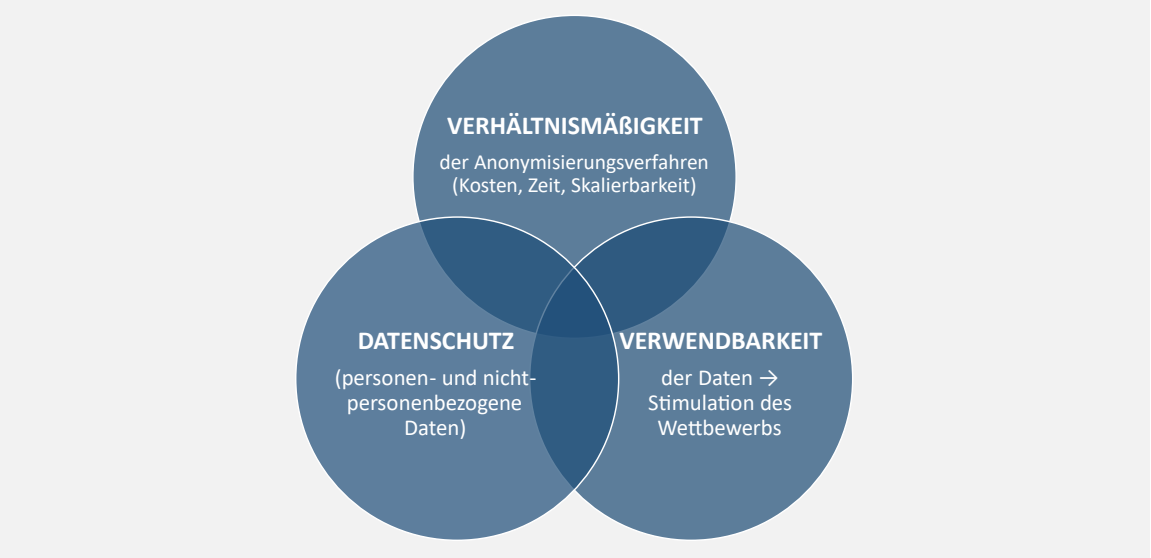
Unternehmen sammeln im Rahmen ihrer Geschäftstätigkeit Daten und sind dabei rechtlich verpflichtet, Datenschutzstandards einzuhalten. Gleichzeitig stehen sie vor der Herausforderung, ihr Geschäftsmodell zu schützen. Dieses führt dazu, dass sie eine möglichst umfassende Anonymisierung bevorzugen, um ihren Wettbewerbsvorteil zu wahren. Diese Praxis kann jedoch als „Sabotage“ wahrgenommen werden, da die Nützlichkeit der Daten durch übermäßige Anonymisierung reduziert wird.[22] Darüber hinaus kann ein hoher Grad an Anonymisierung auch erhebliche Kosten verursachen. Die Implementierung von Anonymisierungsverfahren macht neue Werkzeuge und Technologien erforderlich. Diese Instrumente sind oft mit erheblichen Investitionen in Software und Schulung des beteiligten Personals verbunden.[23]

Auf der anderen Seite stehen Unternehmen, die qualitativ hochwertige und nutzbare Daten benötigen. Sie bevorzugen einen möglichst geringen Anonymisierungsgrad, um die Daten effektiv in ihre Geschäftsprozesse integrieren zu können, was häufig im Konflikt mit den Interessen der Dateninhaber steht.

Die Nutzer von Onlinediensten, die als Datenproduzenten fungieren, erzeugen Daten durch die Nutzung dieser Dienste. Dabei gehen sie eine vertragliche Bindung mit dem Anbieter ein und vertrauen auf den Schutz ihrer persönlichen Daten.

22 Vgl. Stalla-Bourdillon, S./ da Rosa Lazarotto, B. (2024); Stalla-Bourdillon, S./ Knight, A. (2017)
23 Vgl. Farrall, F. et al. (2022).

Abbildung 2-2: Spannungsfelder im Kontext von Datenzugang und Datenaustausch



Quelle: WIK, Eigene Darstellung.

Regulierungsbehörden müssen die Interessen gegeneinander abwägen und zum Ausgleich bringen. Sie werden beurteilen müssen, ob der Dateninhaber unter der Maxime des Datenschutzes mehr anonymisiert als erforderlich ist. Dazu benötigen sie klare Kriterien und Know-how, um Anonymisierungsverfahren zu bewerten und eine Balance zwischen Datenschutz und Datenverfügbarkeit herzustellen.

2.4 Risiko der De-Anonymisierung

Die De-Anonymisierung von Daten beschreibt den gezielten Prozess, anonymisierte Datensätze so zu analysieren, dass die ursprünglich anonymisierten personenbezogenen Informationen wiederhergestellt werden können. Obwohl Anonymisierungstechniken entwickelt wurden, um die Privatsphäre zu schützen, zeigen zahlreiche Studien, dass solche Schutzmaßnahmen häufig unzureichend sind. Dies hat weitreichende Konsequenzen, da die Vorteile der Anonymisierung durch erfolgreiche De-Anonymisierung aufgehoben werden.[24]

Die wachsende Relevanz des Themas ist auf verschiedene Faktoren zurückzuführen. Zum einen führen das exponentielle Wachstum und die allgegenwärtige Generierung von Daten in nahezu allen Lebensbereichen zu einer steigenden Angriffsfläche. Moderne Technologien wie Big Data, Künstliche Intelligenz und Cloud-Computing bieten leistungsfähige Werkzeuge zur Analyse großer Datenmengen. Während diese Technologien ursprünglich für die Datenverarbeitung und -auswertung entwickelt wurden, können sie ebenso dazu verwendet werden, Muster in anonymisierten Datensätzen zu erkennen, die eine Rückverfolgung auf Individuen ermöglichen.[25] Dadurch steigt das Risiko, dass

24 Vgl. Narayanan, A./ Shmatikov, V. (2008).
25 Vgl. Rocher, L. et al. (2019).

sensible Informationen extrahiert und für wirtschaftliche, politische oder sogar kriminelle Zwecke missbraucht werden.

Gleichzeitig erkennen immer mehr Menschen die Bedeutung rechtlicher Vorgaben wie der DSGVO, die Organisationen verpflichten, personenbezogene Daten durch wirksame Anonymisierung zu schützen.[26] Dennoch treten regelmäßig Vorfälle ans Licht, bei denen vermeintlich anonyme Daten nicht ausreichend geschützt waren. Solche Fälle schädigen nicht nur die betroffenen Unternehmen durch rechtliche Konsequenzen, sondern auch das Vertrauen der Öffentlichkeit in den verantwortungsvollen Umgang mit sensiblen Informationen.[27]

Die Bedrohung durch De-Anonymisierung hat zudem erhebliche Auswirkungen auf Innovation und Forschung. In Bereichen wie der Medizin, der Verkehrsplanung oder der Stadtentwicklung ist die Verarbeitung anonymisierter Daten essenziell, um gesellschaftliche und wissenschaftliche Fortschritte zu erzielen. Wenn jedoch das Risiko einer De-Anonymisierung besteht, könnte die Bereitschaft, Daten für solche Zwecke bereitzustellen, erheblich sinken. Dies würde insbesondere Forschungsprojekte behindern, die auf die Analyse großer Datensätze angewiesen sind.

Auch ökonomische Folgen spielen eine entscheidende Rolle. Unternehmen, die von De-Anonymisierungsangriffen betroffen sind, können erhebliche finanzielle Schäden durch Bußgelder, Reputationsverluste oder den Verlust von Kunden erleiden. Der wirtschaftliche Druck, sicherere Anonymisierungstechniken zu entwickeln und sich gegen De-Anonymisierungsrisiken zu wappnen, wächst entsprechend. Die Herausforderung besteht daher darin, den Nutzen von Datenanalysen zu maximieren und gleichzeitig den Schutz der Privatsphäre sicherzustellen.[28]

Verschiedene Techniken und Verfahren können zum Einsatz kommen, um eine De-Anonymisierung durchzuführen. Bereits mehrfach angeklungen ist die Re-Identifikation mittels Verknüpfung. Dabei erfolgt eine Kombination anonymisierter Daten mit externen Datensätzen, die identifizierende Informationen enthalten.[29] Die De-Anonymisierung basierend auf Quasi-Identifikatoren ist ebenfalls bekannt. Dabei werden Attribut wie Postleitzahlen, Geburtsdaten oder Geschlecht genutzt, um Individuen zu identifizieren.[30] Neuere Entwicklungen bieten darüber hinaus Verfahren zur De-Anonymisierung. So gibt es Machine-Learning-basierte Angriffe, die sich den Einsatz von Algorithmen zur Mustererkennung zu Nutze machen. Deep Learning kann verwendet werden, um Verhaltensmuster zu analysieren und Rückschlüsse auf die Identität zu ziehen. Auch die Kombination aus

26 Vgl. Europäisches Parlament (2016).
27 Vgl. Ohm; P. (2010).
28 Vgl. Sweeney, L. (2000).
29 Vgl. Narayanan, A./ Shmatikov, V. (2008).
30 Studien zeigen, dass 87% der US-Bevölkerung anhand von Postleitzahl, Geschlecht und Geburtsdatum eindeutig identifiziert werden können. Vgl. dazu Sweeney, L. (2000). Eine Studie weitere Studie zeigt sogar, dass unter Verwendung von 15 demographischen Attributen bis zu 99,98 % der US-Bevölkerung korrekt in jedem Datensatz identifiziert werden konnten. Vgl. Rocher, L. et al. (2019).

Hintergrundwissen und Wahrscheinlichkeitsmodellen wird von Angreifern genutzt um zu de-anonymisieren.[31]

Die De-Anonymisierung von Daten stellt eine insbesondere im Kontext wachsender Datenmengen und fortschrittlicher Analysetools eine Bedrohung dar. Um den Risiken dieser Bedrohung entgegenzuwirken, ist es für Unternehmen entscheidend, fortschrittliche Anonymisierungsverfahren zu implementieren. Darüber hinaus müssen Anonymisierungspraktiken kontinuierlich überprüft und an neue Angriffsstrategien angepasst werden. Dies sichert nicht nur den Schutz der Privatsphäre, sondern trägt auch dazu bei, das Vertrauen der Öffentlichkeit in die Datensicherheit zu bewahren und eine Balance zwischen Datennutzung und Datenschutz zu gewährleisten.

31 Vgl. El Emam, K. et al. (2011).

3 Daten

3.1 Bestandteile eines Datensatzes und Nutzbarkeit von Daten

Ein Datensatz kann aus mehreren unterschiedlichen Komponenten bestehen, die jeweils eine unterschiedliche Funktion bei der Identifikation und dem Schutz von personenbezogenen Daten haben. Zu den wichtigsten Bestandteilen eines Datensatzes gehören Identifikatoren, Quasi-Identifikatoren und sensible Attribute.[32]

Identifikatoren sind eindeutige Merkmale, die eine Person direkt identifizieren können. Beispiele für Identifikatoren sind der Name, die Ausweisnummer, die Matrikelnummer oder die Sozialversicherungsnummer. Diese Informationen sind so spezifisch, dass sie ohne weitere Hilfsmittel ausreichen, um eine Person eindeutig zu bestimmen. Quasi-Identifikatoren können allein nicht zur eindeutigen Identifikation einer Person verwendet werden, ermöglichen dies jedoch in Kombination mit anderen Attributen. Beispiele für Quasi-Identifikatoren sind Geburtsdatum, Geschlecht, Postleitzahl oder Beruf. Während jedes dieser Merkmale für sich genommen unspezifisch ist, kann eine Kombination mehrerer solcher Merkmale zu einer eindeutigen Identifizierung führen. Sensible Attribute beziehen sich auf schützenswerte persönliche Daten, deren unbefugte Offenlegung zu erheblichen Nachteilen für die betroffene Person führen kann. Beispiele hierfür sind Informationen über das Einkommen, den Gesundheitszustand, ethnische Herkunft, religiöse Überzeugungen oder strafrechtliche Verurteilungen.

Daten können anhand ihrer Sensibilität in verschiedene Kategorien eingeteilt werden, die von der öffentlichen Verfügbarkeit bis hin zur höchsten Schutzwürdigkeit reichen. Daten mit niedriger Sensibilität umfassen allgemein zugängliche Informationen, die öffentlich verfügbar und für den freien Gebrauch vorgesehen sind. Daten mittlerer Sensibilität umfassen interne und vertrauliche Informationen, die in einem geschützten Unternehmensumfeld oder zwischen spezifischen Akteuren geteilt werden. Hochsensible Daten lassen sich in die drei Unterkategorien personenbezogene Daten, sensible personenbezogene Daten und kritische hochsensible Daten unterteilen.

Personenbezogene Daten umfassen gemäß Artikel 4 DSGVO z.B. Kennungen wie einen Namen, eine Kennnummer, oder Standortdaten, da sie Rückschlüsse auf individuelle Personen ermöglichen. Sensible personenbezogene Daten umfassen Informationen über die ethnische Herkunft, denn Gesundheitszustand, genetische und biometrische Daten sowie politische oder religiöse Überzeugungen. Diese Datenkategorie unterliegt gemäß Artikel 9 DSGVO strengeren gesetzlichen Anforderungen, da eine missbräuchliche Verwendung schwerwiegende Folgen haben kann.[33]

---

32 Vgl. Majeed, A./ Lee, S. (2020).

33 Vgl. Europäisches Parlament (2016).

Kritische und hochsensible Daten beinhalten nationale Sicherheitsinformationen, umfangreiche finanzielle Datenbanken sowie Zugangsdaten zu IT-Systemen. Diese Daten gelten als strategisch entscheidend für Unternehmen und Staaten und sind daher von besonderer Bedeutung für Sicherheits- und Compliance-Maßnahmen.

Die Nutzbarkeit von Daten ist für Unternehmen von entscheidender Bedeutung und hängt maßgeblich von der Datenqualität ab. Datenqualität umfasst verschiedene Dimensionen, die sicherstellen, dass Daten den spezifischen Anforderungen der Nutzer entsprechen. Vollständigkeit ist das Ausmaß, in dem alle erforderlichen Daten vorhanden sind. Fehlende Daten können die Analyse erheblich beeinträchtigen. Korrektheit drückt die Übereinstimmung der Daten mit der Realität oder einem Referenzwert aus. Aktualität beurteilt die Angemessenheit des Alters der Daten für die jeweilige Aufgabe. Aktuelle Daten sind oft entscheidend für zeitnahe Entscheidungen. Konsistenz macht Aussagen dazu, inwiefern die Daten in einem einheitlichen Format vorliegen und zu früheren Daten kompatibel sind. Zugänglichkeit ist der Grad, zu dem die Informationen für den Nutzer leicht und schnell abrufbar und nutzbar sind.[34]

Ein weiterer wesentlicher Aspekt der Datenverwertbarkeit ist das Zusammenspiel von Datenqualität und -quantität. In der Regel führt ein Anstieg der Datenmenge zu einer größeren Informationsdichte, die es Unternehmen ermöglicht, tiefere Einblicke zu gewinnen. Besonders im Kontext von Big Data eröffnen sich durch die Analyse umfangreicher Datensätze neue Möglichkeiten zur Mustererkennung und vorausschauenden Analytik. Allerdings bringt die Verarbeitung großer Datenmengen auch erhebliche Herausforderungen mit sich. Unternehmen müssen oft beträchtliche Ressourcen in die Datenintegration, -bereinigung und -analyse investieren, um die potenziellen Vorteile zu realisieren. Die bloße Verfügbarkeit großer Datenmengen stellt per se keinen Nutzen dar. Vielmehr sind der Kontext und die Qualität der Daten entscheidend für den Erfolg der Analysen.[35]

3.2 Typologisierung von Daten

Daten lassen sich auf verschiedene Art typologisieren, was für ihre Analyse und Verarbeitung von zentraler Bedeutung ist. Für den weiteren Verlauf dieser Untersuchung erscheint es angebracht, eine Differenzierung zwischen strukturierten und unstrukturierten Daten vorzunehmen, wie sie beispielsweise in der Informatik üblich ist. Im Anschluss werden die typischen Datenarten großer Onlineplattformen in diese Typologisierung eingeordnet.

Strukturierte Daten sind in einem festen Format organisiert, das eine einfache Verarbeitung und Analyse ermöglicht. Sie finden sich häufig in relationalen Datenbanken und sind durch definierte Felder und Datentypen charakterisiert. Im Gegensatz dazu sind unstrukturierte Daten nicht in einem vorgegebenen Schema organisiert, was ihre Analyse

34 Vgl. Cichy, C./ Rass, S. (2019); Wang, R.Y./ Strong, D.M. (1996).
35 Vgl. Arnold, R. et al. (2020).

komplexer macht. Diese Kategorie umfasst beispielsweise Textdokumente, Bilder und Videos. Innerhalb der unstrukturierten Daten lassen sich zudem repetitive und nicht-repetitive Daten unterscheiden. Repetitive unstrukturierte Daten sind solche, die sich in ähnlicher Form wiederholen, wie etwa E-Mails oder Kundenbewertungen, während nicht-repetitive unstrukturierte Daten einzigartig sind, wie zum Beispiel individuelle Berichte oder spezifische Forschungsdaten.[36]

Onlineplattformen, insbesondere jene, die im Rahmen des DMA[37] als Gatekeeper eingestuft werden, erfassen eine Vielzahl von Nutzerdaten. Diese Daten werden genutzt, um die angebotenen Dienste zu optimieren, personalisierte Werbung zu schalten und eine individuell angepasste Nutzeransprache zu ermöglichen. Abbildung 3-1 zeigt eine Übersicht der Unternehmen und Dienste, die durch DMA und DSA[38] reguliert werden. Der DMA benennt die sogenannten Gatekeeper. Gatekeeper sind große digitale Plattformen, die aufgrund ihrer Marktstellung einen bedeutenden Einfluss auf den digitalen Binnenmarkt ausüben. Sie stellen die Core Platform Services (CPS) bereit. Der DSA benennt VLOP und VLOSE. Diese Kategorie betrifft Plattformen mit mindestens 45 Millionen aktiven Nutzern pro Monat in der EU. CPS und VLOP/VLOSEs haben eine Schnittmenge und fallen unter beide Regulierungen. Sie haben eine zentrale Rolle in den jeweiligen digitalen Ökosystemen der Unternehmen.

Abbildung 3-1: Benennungen im Rahmen von Digital Markets Act und Digital Services Act

Quelle: WIK, Eigene Darstellung.

Diese gesammelten Daten umfassen zunächst **persönliche und eindeutige Identifikationsdaten** wie Name, E-Mail-Adresse und Telefonnummer, die bei der Registrierung abgefragt werden. Diese Daten sind klar strukturiert und in Datenfeldern organisiert.

36 Vgl. dazu und im Folgenden Salinas, S.O./ Lemus, A.C. (2017).
37 Europäisches Parlament (2022b).
38 Europäisches Parlament (2022c).

Zusätzlich werden oft Benutzernamen und Profilbilder gespeichert, um Profile (in sozialen Netzwerken) zu erstellen. Dies sind unstrukturierte Daten, da sie in Form von Bildern vorliegen, die nicht einem festen Datenformat mit vordefinierten Feldern entsprechen. Für den Fall, das Bezahldienste involviert sind, wie z.B. bei Amazon Store oder Booking, werden ebenfalls Zahlungsinformationen, wie Kreditkarten- oder Bankdaten, erfasst. Hier handelt es sich wiederum um strukturierte Daten, da sie in einem spezifischen Format wie Bankkontonummern oder Kreditkartennummern gespeichert werden.

Darüber hinaus sammeln die Plattformen verschiedene **Geräte- und Verbindungsdaten**, darunter IP-Adressen zur Geolokalisierung. Geräteinformationen, wie Betriebssystem, Modell und Browsertyp werden in der Regel ebenfalls gespeichert, ebenso wie Standortdaten, die über GPS, WLAN oder Mobilfunknetze erfasst werden, die beispielsweise bei der Nutzung von Kartendiensten generiert werden. Betriebssystem, Modell und Browsertyp sind strukturierte Daten, die durch definierte Felder und Formate beschrieben werden. GPS-Daten und andere geolokalisierte Daten, sofern sie in Form von Koordinaten oder ähnlichen strukturierten Informationen erfasst werden, gehören ebenfalls zu den strukturierten Daten.

Ein zentraler Bereich sind zudem **Interaktions- und Verhaltensdaten**, mit denen analysiert wird, wie Nutzer auf der Plattform navigieren und welche Inhalte sie in welchem Umfang ansehen. Dazu gehören insbesondere Daten zum Klickverhalten und zur Suchhistorie. Auf Basis dieser Daten wird eine gezielte Personalisierung möglich, da die Interessen der Nutzer ablesbar sind. Auch Kommunikationsdaten, wie Zeitstempel und Metadaten von Direktnachrichten, Kommentaren und „Likes“, werden gesammelt, ohne den Inhalt selbst zu speichern, um basierend darauf Interaktionsmuster und Präferenzen analysieren zu können. Zeitstempel und Metadaten von Direktnachrichten, Kommentaren und „Likes“ sind teilweise unstrukturierte Daten, besonders wenn es um den Inhalt von Nachrichten oder Interaktionen geht. Diese sind nicht durch vordefinierte Datenfelder beschrieben, sondern bestehen aus Text und Interaktionsinformationen. Diese Daten, wie z. B. Kommentare auf sozialen Medien oder Bewertungen von Produkten, können als repetitive unstrukturierte Daten betrachtet werden, da sie regelmäßig in ähnlicher Form auftreten, jedoch in einem unstrukturierten Format.

Kauf- und Transaktionsdaten spielen darüber hinaus eine wichtige Rolle bei E-Commerce-Plattformen oder Diensten mit integrierten Kaufoptionen. Kaufverläufe und Warenkorbdaten werden ebenso gespeichert wie bevorzugte Zahlungsmethoden und Abrechnungsinformationen. Auch dieses dient wiederum dem Zweck, gezielt Angebote unterbreiten zu können und eine personalisierte Nutzererfahrung zu bieten. Kaufverläufe, Warenkorbdaten, bevorzugte Zahlungsmethoden und Abrechnungsinformationen sind strukturiert, da sie klar definierte Felder für Artikel, Preise, Zeitstempel und Zahlungsmethoden enthalten.

Drüber hinaus werden **Sensor- und Kontextdaten** erfasst. Zu ihnen zählen beispielsweise Bewegungs- und Aktivitätsdaten, die durch Bewegungssensoren in Fitness-Apps

oder Augmented-Reality-Diensten erhoben werden. Diese Art von Sensor- oder Kontextdaten kann als unstrukturiert betrachtet werden, wenn sie in Form von Rohdaten von Bewegungssensoren oder Aktivitäts-Tracking-Apps vorliegen, die nicht direkt in einer Datenbank mit vordefinierten Feldern erfasst werden. In bestimmten Anwendungen und Diensten kann ebenfalls auf Mikrofon- und Kameradaten zugegriffen werden, jedoch meist mit Zustimmung der Nutzer.

Auf sozialen Netzwerken werden **Verbindungs- und Netzwerkdaten** gespeichert, die für die Empfehlung und Personalisierung von Inhalten genutzt werden. Zu diesen Daten gehören Informationen zu Freunden und Verbindungen sowie zu Gruppen und Seiten, denen Nutzer folgen, um das Profil und die Interessen der Nutzer zu verfeinern. Diese Daten können sowohl strukturierte als auch unstrukturierte Elemente enthalten. Verbindungen und Freunde können in strukturierten Daten gespeichert werden, aber die Inhalte von Interaktionen oder der Aufbau von Netzwerken (wie Posts und Kommentare) sind meist unstrukturiert.

Diese Liste ist nicht abschließend, gleichwohl stellt sie die wichtigsten Daten, die erhoben werden, dar. Hinzu kommt, dass Unternehmen, die mehrere Dienste betreiben, oft auch **plattformübergreifende Daten**, d.h. Daten, die die Nutzung mehrerer Dienste oder Plattformen des gleichen Unternehmens betreffen, erfassen, um damit umfassendere Nutzerprofile zu erstellen. Auch ist bekannt, dass Daten zu Werbeinteraktionen gespeichert werden. Dazu zählt, dass erfasst wird, welche Anzeigen ein Nutzer gesehen, angeklickt oder gemieden hat, um weitere Anzeigen wiederum besser personalisieren zu können. Daten über Anzeigen, die ein Nutzer gesehen, angeklickt oder gemieden hat, gehören ebenfalls zu repetitiven unstrukturierten Daten, da diese Interaktionen über Zeit in ähnlicher Form wiederkehren.

Die Betrachtung der Daten großer Online-Plattformen zeigt, dass für die weitere Diskussion insbesondere strukturierte und unstrukturiert repetitive Daten von Interesse sind. Gleichwohl können alle drei Arten von Daten im betrachteten Kontext auftreten.

4 Verfahren zur Anonymisierung von Daten

Das nachstehende Kapitel dient dazu, verschiedene Anonymisierungsverfahren systematisch darzustellen, um darauf aufbauend deren Eignung diskutieren zu können. Sie unterscheiden sich zum Teil erheblich in Hinblick auf Aufwand, Nutzen und Compliance.

4.1 Grundlegende Ansätze

Zu den grundlegenden Techniken der Anonymisierung zählen das Entfernen von Quasi-Identifikatoren (auch „Suppression“ genannt), die Datenmaskierung und die Aggregation.

4.1.1 Suppression

Die Suppression bezieht sich auf den Prozess der Löschung oder Verallgemeinerung von Datenattributen, die für sich genommen keine eindeutige Identifizierung ermöglichen, jedoch in Kombination mit anderen Informationen dazu führen können, dass Rückschlüsse auf eine bestimmte Person gezogen werden. Zu den typischen Attributen, die einer Suppression unterzogen werden, gehören beispielsweise Postleitzahl, Geburtsdatum und Geschlecht. Durch das Entfernen oder die Verallgemeinerung dieser Merkmale wird das Risiko einer möglichen Identifizierung verringert.[39]

Die nachfolgende Tabelle 4-1 veranschaulicht verschiedene Beispiele für Suppression. In Zeile 2 wurden sowohl der Name als auch die Postleitzahl entfernt, um eine Identifizierung zu verhindern. In Zeile 3 wurde lediglich die Postleitzahl entfernt, da diese als kritischer Faktor für die Anonymität gilt (teilweise Suppression). In Zeile 4 wurde die Diagnose unterdrückt, da diese in einem spezifischen Kontext als besonders sensibel eingestuft wird (kontextabhängige Suppression).

Tabelle 4-1: Beispiel für eine Suppression

ID	Name	Geburtsdatum	Postleitzahl	Diagnose
1	Max Muster	12.02.1978	12345	Diabetes
2	[unterdrückt]	23.05.1990	[unterdrückt]	Herzflimmern
3	Maria Müller	15.10.1954	[unterdrückt]	Bluthochruck
4	[unterdrückt]	01.01.2011	54321	[unterdrückt]

Quelle: WIK, Eigene Darstellung.

Die Suppression ist besonders vorteilhaft, wenn bestimmte Attribute für den analytischen Zweck nicht unbedingt erforderlich sind und ohne erheblichen Informationsverlust entfernt werden können. Darüber hinaus trägt die Suppression zur Reduktion des Risikos von Datenlecks oder Missbrauch bei, da identifizierende Daten nicht mehr im Datensatz enthalten sind. Diese Technik wird häufig mit dem Konzept der K-Anonymität kombiniert,

39 Vgl. Sweeney, L. (2002a); Pfitzmann, A./ Hansen, M. (2010).

auf das in Kapitel 4.7 näher eingegangen wird. Dieser Ansatz gewährleistet, dass jede Kombination von identifizierbaren Attributen mindestens k Personen im Datensatz betrifft, wodurch die Identifizierbarkeit einzelner Personen deutlich erschwert wird.[40]

4.1.2 Datenmaskierung

Ein weiteres wichtiges Verfahren ist die Datenmaskierung, bei der sensible Daten durch zufällige Zeichen oder Platzhalter ersetzt werden, sodass die ursprünglichen Werte nicht mehr sichtbar sind. Durch die Maskierung bleiben die Daten weiterhin für analytische Zwecke nutzbar, ohne dass die tatsächlichen Werte offengelegt werden. Diese Technik kommt insbesondere in Test- und Entwicklungsumgebungen zum Einsatz, in denen realistische Daten erforderlich sind, gleichzeitig jedoch das Risiko einer Offenlegung von Identitäten oder vertraulichen Informationen minimiert werden soll.[41]

Ein Beispiel für die Anwendung der Datenmaskierung ist die Verarbeitung von Kundendaten für Analyse- oder Testzwecke, wie in der nachstehenden Tabelle 4-2 in zwei Varianten dargestellt. Der Originaldatensatz enthält echte personenbezogene Informationen wie Name, Kreditkartennummer, Adresse und E-Mail. Im ersten maskierten Datensatz werden diese Informationen durch plausible, aber fiktive Werte ersetzt, sodass die Struktur und der Inhalt der Daten erhalten bleiben, jedoch keine Rückschlüsse auf echte Personen möglich sind. Im zweiten Datensatz werden die sensiblen Informationen durch allgemeine Platzhalter ersetzt, die keine realen Daten widerspiegeln, jedoch die Notwendigkeit der Struktur beibehalten, um Tests und Analysen durchzuführen.

Tabelle 4-2: Beispiel für eine Datenmaskierung

Originalformat	Maskierter Datensatz 1	Maskierter Datensatz 2
Name: Max Mustermann	Name: Anna Musterfrau	Name: [NAME]
Kreditkartennummer: 1234 5678 9876 5432	Kreditkartennummer: 4000 1234 5678 9876	Kreditkartennummer: [KREDITKARTENNUMMER]
Adresse: Musterstraße 1, 12345 Musterstadt	Adresse: Beispielstraße 10, 54321 Beispielstadt	Adresse: [ADRESSE]
E-Mail: max.mustermann@email.com	E-Mail: anna.musterfrau@email.com	E-Mail: [E-MAIL]

Quelle: WIK, Eigene Darstellung.

Ein wesentlicher Vorteil der Datenmaskierung besteht in der Erhaltung der Datenintegrität: Trotz der Ersetzung vertraulicher Daten bleibt die Struktur und das Format der ursprünglichen Datensätze erhalten. Dies ermöglicht die weiterhin sinnvolle Nutzung der maskierten Daten für Test- und Analysezwecke, ohne die Funktionalität oder Aussagekraft der Daten zu beeinträchtigen.

40 Vgl. Samarati, P./ Sweeney, L. (1998); Domingo-Ferrer, J./ Torra, V. (2008).
41 Vgl. Samarati, P./ Sweeney, L. (1998); Li, N. et al. (2007).

Ein weiterer wichtiger Vorteil ist die Gewährleistung der rechtlichen Konformität. Durch die Datenmaskierung wird die Einhaltung strenger Datenschutzvorgaben, wie sie durch die DSGVO[42] festgelegt sind, sichergestellt. Die Maskierung stellt sicher, dass keine echten personenbezogenen Informationen in nicht autorisierten Umgebungen verarbeitet werden. Darüber hinaus trägt sie dazu bei, das Vertrauen der Nutzer in die Sicherheit ihrer Daten zu stärken, da die Gefahr einer unbefugten Offenlegung oder Missbrauchs vertraulicher Informationen minimiert wird.

4.1.3 Aggregation

Die Aggregation ist eine Methode, bei der individuelle Datensätze auf eine höhere Ebene der Verallgemeinerung zusammengefasst werden. Anstatt auf spezifische, individuelle Informationen zuzugreifen, werden die Daten so umgestaltet, dass nur aggregierte Ergebnisse wie Durchschnittswerte, Summen oder Häufigkeiten sichtbar sind. Durch diese Umgestaltung wird die Gefahr der Identifikation einzelner Personen erheblich reduziert, da sämtliche Informationen auf Gruppenebene dargestellt werden. Diese Technik wird häufig in Berichten und Analysen eingesetzt, bei denen der Fokus auf der Erkennung von Mustern und Trends liegt, ohne dass personenbezogene Daten auf individueller Ebene zugänglich sein müssen. Aggregation ermöglicht es, wertvolle Erkenntnisse zu gewinnen, während gleichzeitig die Privatsphäre der betroffenen Personen gewahrt bleibt.[43]

Tabelle 4-3: Beispiel für eine Aggregation

Originaldatensatz				
Käufer ID	**Produkt**	**Kategorie**	**Preis (€)**	**Kaufdatum**
A123	Laptop	Elektronik	1.000	12.11.2024
B456	Hose	Kleidung	50	13.11.2024
C789	Buch	Bücher	20	13.11.2024
D123	Smartphone	Elektronik	800	14.11.2024
E456	Jacke	Kleidung	100	14.11.2024
F789	Krimi	Bücher	15	15.11.2024
Anonymisierung mittels Aggregation				
Kategorie	**Anzahl Käufe**	**Durchschnittlicher Preis €**	**Zeitraum**	
Elektronik	2	900	KW 46/2024	
Kleidung	2	75	KW 46/2024	
Bücher	2	17,5	KW 46/2024	

Quelle: WIK, Eigene Darstellung.

Tabelle 4 3 zeigt als Beispiel im oberen Teil die Datensammlung einer Online-Plattform zum Kaufverhalten ihrer Nutzer als Originaldatensatz. Zum Schutz der Privatsphäre der Nutzer erfolgt eine Anonymisierung im unteren Teil der Tabelle. Zu diesem Zweck werden

42 Vgl. Europäisches Parlament (2016)
43 Vgl. Duncan, G. T./ Lambert, D. (1989).

die Kaufdaten nach Produktkategorien gruppiert, und es wird die Anzahl der Käufe pro Kategorie sowie der durchschnittliche Kaufpreis innerhalb eines bestimmten Zeitraums (hier: Kalenderwoche) angegeben. Diese aggregierte Statistik zeigt das Kaufverhalten einer Gruppe von Nutzern, ohne dass Rückschlüsse auf das Verhalten einzelner Nutzer gezogen werden können. Durch die Aggregation werden detaillierte personenbezogene Daten entfernt, wodurch die Identifikation von Einzelpersonen verhindert wird. Gleichzeitig bleiben wichtige Trends und Muster im Kaufverhalten auf Gruppenebene erkennbar. Die Privatsphäre der Nutzer wird geschützt, da individuelle Transaktionsdetails nicht mehr sichtbar sind und eine Re-Identifizierung nicht möglich ist. Die aggregierten Daten, wie durchschnittliche Preise oder Gesamtzahl der Käufe, bleiben jedoch für Marktanalysen, Trendprognosen und die Entwicklung personalisierter Empfehlungen nutzbar. Darüber hinaus unterstützt die Aggregation die rechtliche Konformität, insbesondere im Hinblick auf die DSGVO,[44] da keine direkten personenbezogenen Daten verarbeitet werden.

4.2 Sonderfall: Pseudonymisierung

In der Praxis werden die Begriffe Pseudonymisierung und Anonymisierung oft verwechselt, insbesondere wenn der Personenbezug eines Datensatzes zunächst nicht erkennbar ist. Nach der DSGVO erfordert Anonymisierung die vollständige Entfernung aller identifizierbaren Merkmale, sodass die Daten nicht mehr mit einer natürlichen Person in Verbindung gebracht werden können.[45] Pseudonymisierte Daten gelten jedoch weiterhin als personenbezogen und unterliegen den Datenschutzanforderungen der DSGVO.

Pseudonymisierung beschreibt den Prozess der Ersetzung identifizierender Merkmale durch Pseudonyme, etwa durch die Vergabe von IDs, das Erstellen von Codes oder die Verwendung von Hash-Funktionen und Tokens. Diese Methoden stellen sicher, dass der Personenbezug ohne zusätzliche Informationen nicht unmittelbar erkennbar ist. Im Gegensatz zur Anonymisierung bleibt jedoch eine Re-Identifizierung möglich, wenn die für die Rückführung erforderlichen Informationen, wie ein „Pseudonymisierungsgeheimnis“, vorliegen. Insbesondere Verfahren wie die Tokenisierung, bei der sensible Daten durch bedeutungslose Tokens ersetzt werden, gelten als besonders robust innerhalb der Pseudonymisierung.

Die Anwendung der Pseudonymisierung stellt praktische Herausforderungen dar, vor allem in Bezug auf die sichere Verwaltung des Pseudonymisierungsgeheimnisses. Dieses Geheimnis kann intern durch das datenverarbeitende Unternehmen oder extern durch Dritte verwaltet werden. Eine sorgfältige Analyse des Kontextes ist erforderlich, um Risiken wie Diskriminierung oder Re-Identifizierungsangriffe zu minimieren und gleichzeitig die Nutzbarkeit der Daten zu gewährleisten. Idealerweise sollte das

44 Vgl. Europäisches Parlament (2016).
45 Vgl. Europäisches Parlament (2016)

Pseudonymisierungsgeheimnis in den Händen des datenverarbeitenden Unternehmens verbleiben, das die Datenhoheit besitzt.[46]

Ein Beispiel zur Pseudonymisierung von Kundendaten wird in der nachstehenden Tabelle veranschaulicht. Hierbei wurden Namen und Telefonnummern durch zufällig generierte Pseudonyme ersetzt. Die Kundennummern können jedoch weiterhin einer Person zugeordnet werden, wenn ein zusätzlicher "Schlüssel" (eine separate Tabelle oder Liste) verfügbar ist, der die echten Identitäten enthält.

Tabelle 4-4: Beispiel für eine Pseudonymisierung

Originaldatensatz			
Kundennummer	**Name**	**Adresse**	**Telefonnummer**
201	Max Muster	Musterstr. 1	01234-567890
202	Paula Prototyp	Prototypweg 7	09876-543210
203	Fin Format	Formatallee 102	0115-6789905
Pseudonymisierter Datensatz			
Kundennummer	**Name**	**Adresse**	**Telefonnummer**
201	N1000	Musterstr. 1	T5678
202	N2000	Prototypweg 7	T4321
203	N3000	Formatallee 102	T6789

Quelle: WIK, Eigene Darstellung.

4.3 Noise Addition

Die Technik der Rauschzugabe (Noise Addition) basiert auf der Modifikation der ursprünglichen Daten durch das Hinzufügen zufälliger Werte. Diese zufälligen Veränderungen sorgen dafür, dass die Daten weiterhin für analytische Zwecke genutzt werden können, während jedoch eine exakte Rekonstruktion der Originaldaten ausgeschlossen ist. Ein typisches Anwendungsbeispiel der Rauschzugabe ist die Veröffentlichung von Geodaten, bei der statt der exakten Standorte von Personen geringfügige zufällige Abweichungen eingeführt werden. Dies ermöglicht es, die Datenbasis weiterhin für Analysen, wie etwa zur Erkennung von Trends oder Mustern, zu nutzen, während die Identifizierbarkeit individueller Personen effektiv ausgeschlossen wird. Durch diese Technik wird die Privatsphäre der betroffenen Personen gewahrt, ohne den Wert der Daten für aggregierte Auswertungen und Forschungszwecke zu mindern.[47]

Die Rauschzugabe bietet insbesondere dann Vorteile, wenn bestimmte Attribute im Datensatz potenziell nachteilige Auswirkungen auf Individuen haben könnten. Durch gezielte Modifikationen der Datenattribute wird deren Genauigkeit absichtlich reduziert, während die Gesamtverteilung des Datensatzes erhalten bleibt. Auf diese Weise kann ein externer Betrachter weiterhin von der Genauigkeit der Werte ausgehen, jedoch nur

46 Vgl. European Union Agency for Cybersecurity (ENISA) (2019).
47 Vgl. Mivule, K. (2013).

innerhalb festgelegter Toleranzgrenzen. Bei einer angemessenen Anwendung der Rauschzugabe ist es für eine Drittpartei weder möglich, die Identität der betroffenen Personen zu ermitteln, noch die Daten zu korrigieren oder nachzuvollziehen, wie diese verändert wurden. Diese Technik gewährleistet somit die Anonymität und Datensicherheit, während gleichzeitig die Nützlichkeit der Daten für analytische Zwecke gewahrt bleibt.[48]

Tabelle 4-5 zeigt ein einfaches Beispiel für das Hinzufügen von Rauschen. In diesem Beispiel sollen Mitarbeitergehälter in einem Unternehmen anonymisiert werden, um sicherzustellen, dass die Daten nicht direkt auf Einzelpersonen zurückgeführt werden können. Dazu wird ein zufälliger Wert im Bereich von +/- 500€ zu jedem Gehalt hinzugefügt. Durch diese Modifikation werden die einzelnen Gehaltswerte leicht verzerrt, während der durchschnittliche Trend der Gehaltsdaten weitgehend erhalten bleibt. Diese Methode eignet sich besonders für Szenarien, in denen die Präzision individueller Werte weniger relevant ist als die Analyse von Gesamttendenzen oder Durchschnittswerten. Sie gewährleistet die Anonymität der betroffenen Personen und ermöglicht dennoch aussagekräftige Erkenntnisse.

Tabelle 4-5: Beispiel für eine Noise Addition

ID Mitarbeiter	Gehalt Original (€)	Gehalt mit Noise Addition (€)
1	3.000	3.150
2	4.500	4.300
3	5.000	5.050

Quelle: WIK, Eigene Darstellung.

Das Hinzufügen von Rauschen (Noise Addition) bietet mehrere Vorteile. Durch das gezielte Einfügen zufälliger Abweichungen in die ursprünglichen Daten wird die Privatsphäre der betroffenen Personen geschützt, da individuelle Werte verschleiert werden und Rückschlüsse auf Einzelpersonen erheblich erschwert werden. Gleichzeitig bleibt die Nutzbarkeit der Daten weitgehend erhalten, da die statistischen Eigenschaften wie Mittelwerte oder Verteilungen durch das hinzugefügte Rauschen kaum beeinträchtigt werden. Dies macht die Methode besonders geeignet für Anwendungen wie Machine Learning oder datengetriebene Analysen, bei denen aggregierte Daten erforderlich sind, um Muster zu erkennen oder Vorhersagen zu treffen.

Zusätzlich trägt die Noise Addition zur rechtlichen Konformität bei, insbesondere im Rahmen der DSGVO[49], da sie hilft, personenbezogene Daten in anonymisierte Form zu überführen, die nicht mehr als personenbezogen gelten.

In der Praxis muss die Rauschzugabe häufig mit anderen Anonymisierungstechniken kombiniert werden, wie der Entfernung offensichtlicher Attribute. Das Ausmaß der Rauschzugabe sollte dabei in Abhängigkeit von der erforderlichen

48 Vgl. Independent European advisory body on data protection and privacy (2014).
49 Vgl. Europäisches Parlament (2016).

Informationsgenauigkeit und der Privatsphäre der betroffenen Individuen festgelegt werden, um das Risiko einer Re-Identifikation zu minimieren und die Offenlegung sensibler Attribute zu verhindern.[50]

4.4 Randomisierung

Die Randomisierung ist eine Technik, die Zufallsprozesse nutzt, um Daten zu verändern oder neu anzuordnen, wodurch die Möglichkeit verringert wird, Rückschlüsse auf die ursprünglichen Werte zu ziehen. Durch den Einsatz zufälliger Veränderungen wird der Zusammenhang zwischen den Originaldaten und den veränderten Werten so stark verschleiert, dass die Identifizierung der betroffenen Personen erschwert wird. Diese Methode wird häufig verwendet, um Daten zu anonymisieren, ohne ihre Nutzbarkeit für statistische Auswertungen zu verlieren. Sie ermöglicht es, die Daten in einer Weise zu transformieren, die sie für Analysezwecke weiterhin wertvoll macht, aber gleichzeitig das Risiko der Re-Identifikation minimiert.[51]

Ein häufig verwendetes Beispiel für Randomisierung ist das Hinzufügen von (Zufalls-)Rauschen, indem zu jedem Wert eine zufällige Störung oder Abweichung hinzugefügt wird. Diese Veränderung verhindert, dass eine direkte Verbindung zwischen den anonymisierten Daten und den ursprünglichen Informationen gezogen werden kann, während die Grundmuster und Trends in den Daten erhalten bleiben.

Ein Vorteil der Randomisierung liegt darin, dass sie einerseits die Privatsphäre der betroffenen Personen schützt, indem sie die exakten Werte verändert. Andererseits ermöglicht sie, dass aggregierte statistische Analysen weiterhin durchführbar sind.

Es gibt jedoch auch Herausforderungen und potenzielle Nachteile bei der Anwendung der Randomisierung. Zum einen kann das Hinzufügen von Zufallsrauschen oder das Verändern von Datenpunkten zu einer Verzerrung der Daten führen, was insbesondere in hochsensiblen Anwendungsbereichen problematisch sein kann. Wenn beispielsweise zu viele zufällige Modifikationen an einem Datensatz vorgenommen werden, könnten die Ergebnisse von Analysen weniger genau werden. Zudem erfordert die Auswahl der richtigen Zufallsveränderung ein sorgfältiges Abwägen, um sowohl die Privatsphäre zu wahren als auch die Daten weiterhin für valide Schlussfolgerungen verwenden zu können.

In der Praxis wird Randomisierung oft in Bereichen eingesetzt, in denen genaue Werte für den einzelnen Datenpunkt nicht notwendig sind, aber dennoch ein allgemeines Verständnis oder Muster der Daten erforderlich ist. Beispiele hierfür sind Marktforschung, Gesundheitsstatistiken oder Verhaltensanalysen, bei denen Trends und Muster von Gruppen von Interesse sind, jedoch keine individuellen Details im Vordergrund stehen.

50 Vgl. Majeed, A. (2023).
51 Vgl. Sweeney, L. (2002b); Fung, B.C.M. et al. (2010).

Nachstehende Tabelle 4-6 zeigt das Beispiel einer Anonymisierung von Gehaltsdaten mit einer zufälligen Veränderung im Bereich von +/-500 € bei einer Beibehaltung der statistischen Verteilung, d.h. der Durchschnitt der randomisierten Werte liegt nahe am Originaldatensatz.

Tabelle 4-6: Beispiel für eine Randomisierung

ID Mitarbeiter	Abteilung	Gehalt (€)	Gehalt (€) randomisiert
1	IT	3.500	3.600
2	HR	2.800	2.750
3	Marketing	4.200	4.150
4	Öffentlichkeitsarbeit	3.800	3.900
5	HR	3.000	2.950

Quelle: WIK, Eigene Darstellung.

Die genauen Gehälter sind nicht mehr nachvollziehbar und die Privatsphäre der Betroffenen wird geschützt. Gleichwohl sind durch die Beibehaltung der statistischen Verteilung analytische Auswertungen möglich. Der Mittelwert und andere aggregierte Kennzahlen, wie die Varianz, bleiben weitgehend unverändert. Es besteht eine Balance zwischen Datenschutz und Datenverwendbarkeit.[52]

Insgesamt stellt die Randomisierung eine flexible und weit verbreitete Methode dar, um eine Balance zwischen Datenschutz und der analytischen Nützlichkeit von Daten zu erreichen. Sie ermöglicht es, die Privatsphäre zu wahren, während gleichzeitig wertvolle statistische Informationen für die Analyse und Entscheidungsfindung erhalten bleiben.

4.5 Permutation

Die Permutation ist eine spezielle Form der Randomisierung, bei der die Werte innerhalb eines Datensatzes neu angeordnet oder vertauscht werden, ohne dabei die ursprünglichen Daten zu verändern. Diese Technik bewahrt die Datenintegrität, macht es jedoch deutlich schwieriger, die Daten mit bestimmten Individuen zu verknüpfen. Bei der Permutation bleiben die Originalwerte erhalten, aber ihre Positionen oder Zuordnungen innerhalb des Datensatzes werden verändert. Auf diese Weise wird verhindert, dass bestimmte Datenpunkte einer konkreten Person zugeordnet werden können, was insbesondere im Bereich des Datenschutzes von Bedeutung ist. Ein typisches Beispiel für die Anwendung der Permutation ist die Umordnung von Zeitstempeln oder geographischen Koordinaten in einem Datensatz, sodass die Beziehung zwischen den Daten und den Individuen verschleiert wird.[53]

---

52 Bei sehr präzisen oder sensiblen Analysen können durch die Randomisierung allerdings auch Verzerrungen einzelner Werte verursacht werden. Wenn der Zufallsbereich zu groß gewählt wird, könnten sogar aggregierte Kennzahlen wie Durchschnitt oder Median in ihrer Genauigkeit beeinträchtigt werden, was weniger präzise Ergebnisse zur Folge hätte.

53 Vgl. Bender, A. (2015).

Die folgende Tabelle gibt ein Beispiel für die Anonymisierung von Mitarbeitergehältern in einem Unternehmen. Zur Permutation werden die Werte in der Spalte Gehalt zufällig vertauscht. Im Gegensatz zur Randomisierung werden dabei keine neuen Werte erzeugt, sondern die Reihenfolge der Originalwerte vertauscht.

Tabelle 4-7: Beispiel für eine Permutation

ID Mitarbeiter	Abteilung	Gehalt (€)	Gehalt (€) permutiert
1	IT	3.500	2.800
2	HR	2.800	3.800
3	Marketing	4.200	3.500
4	Öffentlichkeitsarbeit	3.800	3.000
5	HR	3.000	4.200

Quelle: WIK, Eigene Darstellung.

Die Permutation gewährleistet, dass die direkte Identifikation einzelner Personen aus einem Datensatz nicht mehr möglich ist. Gleichzeitig bleibt die Datenintegrität gewahrt, da die grundlegenden Verteilungen und statistischen Merkmale der Daten unangetastet bleiben. Analysemöglichkeiten für aggregierte Auswertungen sind damit möglich. Gleichwohl können unter Umständen Rückschlüsse auf Gruppen oder Subpopulationen möglich sein, was bei der Planung von Datenschutzmaßnahmen zu berücksichtigen ist.

4.6 Data Swapping

Data Swapping ist eine Anonymisierungstechnik, die durch den Tausch von Attributen zwischen Datensätzen die Privatsphäre der Individuen schützt. Ein wesentlicher Vorteil dieser Methode liegt in der Beibehaltung der statistischen Eigenschaften des Datensatzes: Die Häufigkeiten der Werte bleiben unverändert, sodass sich die Daten weiterhin für statistische Analysen eignen. Zudem erschwert Data Swapping die Identifikation einzelner Personen, da sensible Informationen nicht mehr direkt mit bestimmten Individuen verknüpft sind.

Die Methode bringt auch Nachteile mit sich. Bei intensivem Swapping kann es zu einem Informationsverlust kommen, da durch das Verfälschen der Zuordnungen die Genauigkeit des Datensatzes leidet. Bei komplexeren Datensätzen kann es zudem schwierig sein, die Konsistenz zwischen den verschiedenen Attributen zu gewährleisten, wenn mehrere Attribute gleichzeitig vertauscht werden. Ein typisches Anwendungsbeispiel für Data Swapping sind öffentliche Datensätze, wie zum Beispiel demografische Umfragen, bei denen der Schutz der Privatsphäre dadurch verbessert werden kann, ohne die statistische Nutzbarkeit stark einzuschränken.[54]

54 Vgl. Wang, K./ Li, J. (2005).

Dem nachstehenden Beispiel ist die Vorgehensweise für das Swapping zu entnehmen. Die Attribute Alter und Wohnort wurden zwischen den Personen getauscht, ohne die Häufigkeiten dieser Werte im gesamten Datensatz zu verändern.

Tabelle 4-8: Beispiel für Swapping

Orignaldatensatz				
ID Person	**Name**	**Alter**	**Wohnort**	**Beruf**
A	Anna	34	Bonn	Polizist
B	Benjamin	29	Hamburg	Ingenieur
C	Charlotte	42	Stuttgart	Arzt
D	Daniel	38	Regensburg	Notar
Datensatz nach Swapping				
ID Person	**Name**	**Alter**	**Wohnort**	**Beruf**
A	Anna	42	Regensburg	Polizist
B	Benjamin	38	Stuttgart	Ingenieur
C	Charlotte	34	Bonn	Arzt
D	Daniel	29	Hamburg	Notar

Quelle: WIK, Eigene Darstellung.

Damit bleiben die statistischen Eigenschaften des Datensatzes wie z.B. das Durchschnittsalter erhalten. Gleichzeitig ist die direkte Verknüpfung zwischen den Individuen und ihren sensiblen Attributen jedoch aufgebrochen. Ein Nachteil ist, dass die Konsistenz der Daten zwingend gegeben sein muss, indem die Werte der Attribute nicht unbedingt zusammenpassen müssen. Dieses wäre z.B. dann der Fall, wenn Alter oder Gehalt und Beruf nicht konsistent zueinander passen.

4.7 Data Hashing

Das **Hashen** von Daten ist weit verbreitet, insbesondere in Bereichen wie der Speicherung von Passwörtern oder der Sicherstellung der Datenintegrität. Dabei werden sensible Daten durch einen Hash-Algorithmus in einen festen, meist kürzeren Wert umgewandelt, der die ursprünglichen Daten nicht mehr rekonstruierbar macht. Einmal gehashte Daten sind nicht mehr rekonstruierbar, was bedeutet, dass eine Rückgewinnung der ursprünglichen Informationen ohne zusätzliche Mechanismen nicht möglich ist.[55]

Tabelle 4-9 zeigt ein Beispiel für Hashing in dem Benutzername, Passwort und E-Mail anonymisiert werden. Die Original-Daten wurden mit einem Hash-Algorithmus (z. B. SHA-256) verarbeitet. Ein Hash-Algorithmus, wie z. B. SHA-256 (Secure Hash Algorithm 256-Bit), ist eine kryptografische Funktion, die Eingabedaten beliebiger Größe in eine

55 Vgl. Vardalachakis, M.et al. (2023).

fixe, einzigartige Zeichenfolge, den sogenannten Hash-Wert, umwandelt.[56] Das Ergebnis ist ein eindeutiger, nicht rückrechenbarer Wert (Hash).

Tabelle 4-9: Beispiel für Hashing

Originaldaten	Daten gehasht (SHA-256)
Benutzername: Anna_Muster	79ed9532d3aabd4833eee3acc1f66176d413a6c46f597728d00753fa52a6931f
Passwort: P@ssw0rd123	231ecc7d178da5f22983bc579599396d6c139a457987ae1ee0026d88432d6a72
E-Mail: anna@example.com	f7d54c22dda5b1780276601cd005909b70543fc4a6ba72d735b4cb9bbb611f01

Quelle: WIK, Eigene Darstellung unter Verwendung von https://rechneronline.de/hash/sha256.php.

Die Original-Daten können nicht direkt aus dem Hash rekonstruiert werden, was die Privatsphäre schützt. Die gehashte Version kann für Vergleiche verwendet werden (z. B. zur Passwortüberprüfung), ohne die ursprünglichen Daten speichern zu müssen. Dazu wird der resultierende Hash-Wert mit einem bekannten oder erwarteten Wert verglichen. Wenn die Hash-Werte übereinstimmen, ist sichergestellt, dass die Werte seit dem ursprünglichen Hashing nicht verändert worden sind.

Bei Verwendung eines schwachen oder kompromittierten Hash-Algorithmus könnten Angriffe möglich sein. Die Daten sind dann sicher, wenn der Hash einzigartig und die zugrunde liegenden Daten ausreichend komplex sind. Typische Beispiele sind die Speicherung von Passwörtern in Datenbanken und die Anonymisierung eindeutiger Identifikatoren (z. B. Kunden-IDs).[57]

4.8 Generalisierung mittels K-Anonymität, L-Diversität und T-Closeness

Die Generalisierung ist eine Technik der Datenanonymisierung, bei der spezifische Daten durch allgemeinere Kategorien ersetzt werden, um die Identifizierbarkeit von Individuen zu reduzieren. Dieser Ansatz funktioniert, indem detaillierte Informationen in weniger präzise, aber immer noch nützliche Daten umgewandelt werden. So kann beispielsweise das genaue Geburtsdatum einer Person durch eine Altersspanne ersetzt werden. Dadurch werden spezifische Merkmale, die zu einer Identifizierung führen könnten, unkenntlich gemacht, ohne die grundlegenden Daten für Analysen unbrauchbar zu machen. Zu den bekanntesten und am weitesten verbreiteten Techniken der Generalisierung gehören K-Anonymität und L-Diversität.[58]

56 Im Internet werden verschiedene Rechner für eine solche Umrechnung zur Verfügung gestellt. Vgl. z.B. https://rechneronline.de/hash/sha256.php [Letzter Abruf 19.11.2024].
57 Vgl. dazu https://www.ssldragon.com/de/blog/sha-256-algorithmus/#hashing-definition.
58 Vgl. Slijepcevic, D. et al (2021).

4.8.1 K-Anonymität

Das K-Anonymitätsmodell ist eine etablierte Methode der Datenanonymisierung. Der Ansatz basiert darauf, dass mindestens k Personen innerhalb einer Äquivalenzklasse identische Werte für alle relevanten Quasi-Identifikatoren (z. B. Alter, Wohnort) aufweisen müssen. Dies erschwert die Re-Identifikation von Individuen, da diese nur noch als Teil einer Gruppe und nicht mehr als Einzelperson wahrgenommen werden können.[59]

Im Rahmen von K-Anonymität wird ein Datensatz so transformiert, dass jede Kombination von Quasi-Identifikatoren mindestens k-mal vorkommt. Ein Beispiel für k=2 bedeutet, dass mindestens zwei Personen mit denselben Quasi-Identifikatoren vorhanden sein müssen. Dadurch liegt die Wahrscheinlichkeit, eine Person eindeutig zu identifizieren, bei höchstens 1/k. Mit steigenden k-Werten nimmt der Schutz der Privatsphäre zu, jedoch kann dies die Nutzbarkeit der Daten für Analysen einschränken, da die Daten zunehmend generalisiert oder aggregiert werden müssen. Im Folgenden soll in ein einfaches Beispiel zur Illustrierung gegeben werden. [60]

Tabelle 4-10: Beispiel für K-Anonymität (k=3)

Originaldatensatz				
Alter	**Wohnort**	**Geschlecht**	**Diagnose**	
34	Stuttgart	Weiblich	Diabetes	
35	München	Männlich	Diabetes	
36	Regensburg	Weiblich	Bluthochdruck	
34	Augsburg	Männlich	Grippe	
57	Hamburg	Weiblich	Krebs	
58	Kiel	Weiblich	Grippe	
60	Oldenburg	Männlich	Herzinsuffizienz	
Datensatz nach K-Anonymität (k=3)				
Alter	**Wohnort**	**Geschlecht**	**Diagnose**	**Gruppe**
30-40	Süddeutschland	*	Diabetes	1
30-40	Süddeutschland	*	Diabetes	1
30-40	Süddeutschland	*	Bluthochdruck	1
30-40	Süddeutschland	*	Grippe	1
50-60	Norddeutschland	*	Krebs	2
50-60	Norddeutschland	*	Diabetes	2
50-60	Norddeutschland	*	Herzinsuffizienz	2

Quelle: WIK, Eigene Darstellung.

Der Datensatz enthält Patientendaten und umfasst die Attribute Alter, Wohnort, Geschlecht und Diagnose und soll mit der Bedingung k=3 anonymisiert werden. Die Altersangaben wurden in Spannen generalisiert, die Wohnorte aggregiert und das Geschlecht

59 Majeed, A./ Lee, S. (2020); Winter, C. et al. (2019).
60 Sweeney, L. (2002); Iyengar, J. (2002).

unterdrückt. Für jede Gruppe (Gruppe 1 und Gruppe2) wird nun k=3 erfüllt, da mindestens drei Personen denselben Quasi-Identifikator (Alter, Wohnort, Geschlecht) teilen. Dadurch wird eine Re-Identifikation der individuellen Patienten erschwert.

Auch wenn K-Anonymität ein robustes Modell ist, bietet sie keinen vollständigen Schutz vor Re-Identifizierung. So enthält Gruppe 1 zwei gleiche Wert für die Diagnose und der Schutz des sensiblen Attributs ist eingeschränkt. Wenn ein Angreifer Zugang zu weiteren externen Datenquellen (Zusatzinformationen) hat, kann die Anonymisierung umgangen werden. Auch können sensible Attribute (z. B. Krankheiten) innerhalb einer Äquivalenzklasse für alle Datensätze gleich sein, wodurch ihre Offenlegung möglich wird, obwohl die Quasi-Identifikatoren anonymisiert sind.

Die Wahl des k-Werts ist ein entscheidender Faktor für die Vorteilhaftigkeit der Methode. Ein niedriger Wert birgt das Risiko unzureichender Anonymität, während ein hoher Wert die Daten für Analysen stark verfälschen kann. Daher ist es wichtig, den k-Wert auf die spezifischen Datenschutzanforderungen und Anwendungsziele abzustimmen.

4.8.2 L-Diversität

Das L-Diversitäts-Modell ist eine Erweiterung des K-Anonymitätsmodells und wurde entwickelt, um die Schwächen von K-Anonymität beim Schutz sensibler Attribute zu adressieren. Während K-Anonymität sicherstellt, dass jede Äquivalenzklasse mindestens k Datensätze umfasst, geht L-Diversität einen Schritt weiter: Sie garantiert, dass innerhalb jeder Äquivalenzklasse mindestens l unterschiedliche, gut repräsentierte Werte für das sensible Attribut vorhanden sind. Dadurch wird das Risiko verringert, dass eine Einzelperson anhand der Werte eines sensiblen Attributs re-identifiziert oder dass deren Informationen offengelegt werden können.

Eine Äquivalenzklasse erfüllt die Anforderungen von L-Diversität, wenn die Werte des sensiblen Attributs ausreichend divers sind. Beispielsweise sorgt eine 3-Diversität dafür, dass mindestens drei unterschiedliche Ausprägungen für das sensible Attribut in jeder Gruppe vorkommen. Dies erschwert es, präzise Rückschlüsse auf einzelne Personen zu ziehen oder ihre sensiblen Informationen zu bestimmen. Das bekannte Beispiel zur K-Anonymität dient hier der Illustration und wird in Tabelle 4-11 fortgesetzt.

Alter und Wohnort wurden hier, wie bei der K-Anonymität auch, aggregiert. Die L-Diversität erfordert, dass jede Gruppe mindestens drei unterschiedliche Diagnosen (Ausprägungen des sensiblen Attributs) enthält. Es zeigt sich, dass in diesem Beispiel der gleiche Datensatz, der das Kriterium für eine K-Anonymität von drei erfüllt, auch das Kriterium eine L-Diversität von drei erfüllt.

Tabelle 4-11: Beispiel für L-Diversität (l=3)

Originaldatensatz				
Alter	**Wohnort**	**Geschlecht**	**Diagnose**	
34	Stuttgart	Weiblich	Diabetes	
35	München	Männlich	Diabetes	
36	Regensburg	Weiblich	Bluthochdruck	
34	Augsburg	Männlich	Grippe	
57	Hamburg	Weiblich	Krebs	
58	Kiel	Weiblich	Grippe	
60	Oldenburg	Männlich	Herzinsuffizienz	
Datensatz nach L-Diversität (l=3)				
Alter	**Wohnort**	**Geschlecht**	**Diagnose**	**Gruppe**
30-40	Süddeutschland	*	Diabetes	
30-40	Süddeutschland	*	Diabetes	1
30-40	Süddeutschland	*	Bluthochdruck	1
30-40	Süddeutschland	*	Grippe	1
50-60	Norddeutschland	*	Krebs	2
50-60	Norddeutschland	*	Grippe	2
50-60	Norddeutschland	*	Herzinsuffizienz	2

Quelle: WIK, Eigene Darstellung.

L-Diversität ist eine wirksame Technik zur Erhöhung des Datenschutzes, die speziell den Schutz sensibler Attribute im Fokus hat. An ihre Grenzen stößt sie dann, wenn ein Wert für das sensible Attribut stark dominiert. In solchen Fällen bleibt das Risiko bestehen, dass durch Wahrscheinlichkeit Rückschlüsse mit hoher Genauigkeit gezogen werden können. Das Modell berücksichtigt bei der Anonymisierung nicht die Verteilung und Ähnlichkeit der Quasi-Identifikatoren. Dadurch kann die Nutzbarkeit der anonymisierten Daten beeinträchtigt werden, insbesondere wenn zu viele Einschränkungen eingeführt werden müssen, um Diversität zu erreichen.[61]

4.8.3 T-Closeness

Um die Schwächen von K-Anonymität und L-Diversität zu beheben, ist das T-Closeness-Modell zur Anonymisierung tabellarischer Daten entwickelt worden. Das Modell fokussiert auf den Schutz der Verteilung der sensiblen Attribute innerhalb einer Äquivalenzklasse.

Eine Tabelle erfüllt dann T-Closeness, wenn die Verteilung des sensiblen Attributs in den Äquivalenzklassen höchstens um eine Distanz t von der Gesamtverteilung in der gesamten Tabelle abweicht. Der Schwellenwert t wird je nach gewünschtem Datenschutz und

61 Vgl. Machanavajjhala, A et al. (2007).

Zweck der Datenveröffentlichung festgelegt. Für das bekannte Beispiel ist diese Bedingung so nicht erfüllbar.[62]

Im Folgenden soll das Modell anhand des bekannten Beispiels in Tabelle 4-12 erläutert werden. Es soll t=0,2 gelten, d.h. die Verteilung des sensiblen Attributs in der Gruppe darf maximal 20% von der Verteilung im Originaldatensatz abweichen. Eine derartige Verteilung kann mit der vorhandenen geringen Zahl an Fällen bzw. Datensätzen nicht erreicht werden. Vor diesem Hintergrund ist es erforderlich, synthetisch einzugreifen und den Datensatz um einige Fälle zu ergänzen.

Auch hier wurden die Altersgruppen und Regionen zusammengefasst und das

Tabelle 4-12: Beispiel T-Closeness (t=0,2)

Originaldatensatz					
Alter	**Wohnort**	**Geschlecht**	**Diagnose**		**Verteilung**
34	Stuttgart	Weiblich	Diabetes		28,6
35	München	Männlich	Diabetes		28,6
36	Regensburg	Weiblich	Bluthochdruck		14,3
34	Augsburg	Männlich	Grippe		28,6
57	Hamburg	Weiblich	Krebs		14,3
58	Kiel	Weiblich	Grippe		28,6
60	Oldenburg	Männlich	Herzinsuffizienz		14,3
Datensatz nach T-Closeness (t=0,2)					
Alter	**Wohnort**	**Geschlecht**	**Diagnose**	**Gruppe**	**Verteilung**
30-40	Süddeutschland	*	Diabetes	1	33,3%
30-40	Süddeutschland	*	Diabetes	1	33,3%
30-40	Süddeutschland	*	Bluthochdruck	1	16,7%
30-40	Süddeutschland	*	Grippe	1	33,3%
30-40	Süddeutschland	*	Herzinsuffizienz (synth.)	1	16,7%
30-40	Süddeutschland	*	Grippe (synth.)	1	33,3%
50-60	Norddeutschland	*	Krebs	2	16,7%
50-60	Norddeutschland	*	Grippe	2	33,3%
50-60	Norddeutschland	*	Herzinsuffizienz	2	16,7%
50-60	Norddeutschland	*	Grippe (synth.)	2	33,3%
50-60	Norddeutschland	*	Diabetes (synth.)	2	33,3%
50-60	Norddeutschland	*	Diabetes (synth.)	2	33,3%

Quelle: WIK, Eigene Darstellung.

Geschlecht durch einen Platzhalter ersetzt. Für die gebildeten Gruppen gilt nun eine ähnliche Verteilung wie im Originaldatensatz. Im Originaldatensatz haben die Diagnosen Diabetes und Grippe eine Verteilung von 2/7=28,6% und Bluthochdruck, Krebs und

62 Vgl. Li, N. et al. (2007).

Herzinsuffizienz von 1/7=14,3%. Dieses bedeutet, dass bei einer zulässigen Abweichung von 20% eine Verteilung im Bereich von 28,6*1,2=34,32% bis 28,6*0,8=22,88% für Diabetes und Grippe zulässig ist und eine Verteilung im Bereich von 14,3*1,2=17,16% bis 14,3*0,8=11,44% für Bluthochdruck, Krebs und Herzinsuffizienz. Diese Bedingungen sind in beiden Gruppen nach dem Hinzufügen von synthetischen Daten erfüllt.

Das T-Closeness-Modell verbessert den Schutz sensibler Daten, kann jedoch die Nutzbarkeit der anonymisierten Daten deutlich verringern.

4.8.4 Vergleich der Modelle

Die drei Modelle K-Anonymität, L-Diversität und T-Closeness beinhalten unterschiedliche Ansätze zum Schutz der Privatsphäre und zur Datenanonymisierung, die je nach Anwendungsfall ihre Stärken und Herausforderungen aufweisen.

K-Anonymität stößt in der Praxis insbesondere in Bezug auf Skalierbarkeit und Kosteneffizienz bei großen Datenmengen durchaus an ihre Grenzen. Der technische Aufwand, um sicherzustellen, dass jeder Datenpunkt in einer Gruppe von mindestens k ununterscheidbar bleibt, wächst exponentiell, wenn die Datenmenge oder der Wert von k steigt. Dennoch kann K-Anonymität mit gut optimierten Algorithmen und automatisierten Anonymisierungswerkzeugen effizient eingesetzt werden. Eine skalierbare Lösung erfordert eine Infrastruktur, die auch mit wachsenden Datenmengen umgehen kann, ohne die Betriebskosten unverhältnismäßig zu steigern. L-Diversität reduziert die Wahrscheinlichkeit von Rückschlüssen auf Einzelpersonen und bietet somit einen stärkeren Datenschutz. Jedoch besteht oft die Notwendigkeit, mehr Werte in jede Gruppe zu integrieren was wiederum einen höheren technischen Aufwand und damit zu steigenden Kosten. Insbesondere bei großen Datenmengen und hohen Werten von l wird die Skalierbarkeit zu einer Herausforderung. Automatisierte Prozesse und effiziente Algorithmen sind auch hier notwendig, um die Implementierung effizient und kostengünstig zu gestalten. T-Closeness geht noch einen Schritt weiter, indem sie sicherstellt, dass die Verteilung sensitiver Attribute innerhalb einer Gruppe der Gesamtverteilung der Originaldaten ähnelt. Dies bietet eine zusätzliche Schutzebene, kann jedoch bei großen Datensätzen, die viele Verteilungsberechnungen erfordern, den technischen Aufwand und die Kosten erhöhen. Die Skalierbarkeit von T-Closeness ist von der Komplexität der Daten und der Anzahl der sensitiven Attribute abhängig. Der Einsatz von optimierten Algorithmen und automatisierten Prozessen ist entscheidend.

Zusammenfassend lässt sich sagen, dass alle drei Modelle einen höheren Datenschutz bieten, jedoch mit unterschiedlichen Herausforderungen hinsichtlich der Skalierbarkeit und Kosteneffizienz. T-Closeness und L-Diversität erfordern mehr Berechnungen und daher auch mehr technische Ressourcen als K-Anonymität. In großen und komplexen Datensätzen müssen alle drei Modelle durch optimierte und automatisierte Prozesse

unterstützt werden, um sowohl ihre Skalierbarkeit als auch ihre Kosteneffizienz zu gewährleisten.

Abschließend sollen noch ein paar Aussagen zur Wahl der Parameter gemacht werden. Die Wahl des Parameters k bei K-Anonymität und l bei L-Diversität hängt von der Art und Sensibilität der Daten sowie den angestrebten Datenschutz – und Nutzbarkeitszielen ab. Es gibt jedoch allgemeine Empfehlungen und Überlegungen, die hier kurz dargelegt werden sollen.

Bei der K-Anonymität gilt, dass in der Praxis häufig ein Wert für k zwischen 3 und 10 angesetzt wird. Ein höheres k, von z.B. 20, wird für besonders sensible Daten empfohlen, da es die Wahrscheinlichkeit einer eindeutigen Identifizierung einer Person stark reduziert. Dazu bedarf es aber auch entsprechend großer Datensätze. Ein niedriger k-Wert von z.B. 2 bietet nur minimalen Schutz, da Gruppen mit nur wenigen Personen immer noch eine erhöhte Rückverfolgbarkeit haben. Ein sehr hoher k-Wert kann aber auf der anderen Seite die Datenintegrität und Nutzbarkeit stark einschränken, da viele Details aggregiert werden müssen.[63]

Bei der L-Diversität liegen übliche Werte für l im Bereich von 3 bis 5, abhängig von der Verteilung der sensiblen Attribute. Für hochsensible Daten kann der Wert von l bei 10 oder mehr liegen. L-Diversität ist dann besonders gut geeignet, wenn die sensiblen Attribute eine breite Verteilung aufweisen. Ein hoher Wert ist wichtig, wenn eine Korrelation zwischen den quasi-identifizierenden Attributen und sensiblen Daten besteht, um Angriffe zu verhindern.[64]

Für den Parameter t gibt es keine festen, universellen Werte. T-Werte werden häufig basierend auf der Sensibilität der Daten und den Sicherheitsanforderungen angepasst. In der Praxis empfehlen viele Studien, dass der Wert von t klein gehalten wird, um einen signifikanten Schutz gegen Re-Identifikationsangriffe zu gewährleisten. Ein Bereich von t=0,1 bis 0,2 wird oft als ausreichend angesehen, um einen guten Schutz zu bieten, ohne die Daten zu stark zu verzerren. Ein kleinerer Wert bietet besseren Schutz, könnte aber die Nutzbarkeit der Daten beeinträchtigen. In einigen Szenarien kann ein Wert von t=0,5 ausreichen, insbesondere bei weniger sensiblen Daten. Dieses bedeutet jedoch einen größeren Kompromiss zwischen Schutz und Nutzbarkeit der Daten.[65]

4.9 Differential Privacy

Differential Privacy ist eine Methode, die in der Datenanalyse und im maschinellen Lernen verwendet wird, um den Schutz der Privatsphäre in Datensätzen zu gewährleisten. Sie bietet eine mathematische Garantie, dass die Privatsphäre einzelner Personen gewahrt bleibt, selbst wenn statistische Analysen auf den Daten durchgeführt werden.

63 Vgl. Sweeney, L. (2002b); Iyengar, J. (2002).
64 Vgl. Machanavajjhala, A et al. (2007).
65 Vgl. Li, N. et al. (2007).

Durch Techniken wie das Hinzufügen von Rauschen wird gewährleistet, dass aggregierte Erkenntnisse über eine Gruppe gewonnen werden können, ohne Informationen über einzelne Individuen preiszugeben.[66]

Das Grundprinzip der Differential Privacy besteht darin, dass die Ergebnisse einer Datenanalyse unabhängig davon sind, ob die Daten einer bestimmten Person im Datensatz enthalten sind oder nicht. Formal bedeutet dies, dass für zwei Datensätze, die sich nur in den Daten einer einzigen Person unterscheiden, die Wahrscheinlichkeit, dass ein Algorithmus einen bestimmten Wert liefert, nahezu gleichbleibt.[67] Um dieses Ziel zu erreichen, wird typischerweise Laplace-Rauschen hinzugefügt, das auf der Laplace-Verteilung basiert und entsprechend dem Parameter ε (Epsilon) skaliert wird. Dieser Parameter steuert das Gleichgewicht zwischen Datenschutz und Genauigkeit der Ergebnisse.

Differential Privacy hat sich als De-facto-Standard für den Schutz der Privatsphäre durchgesetzt und wird von zahlreichen Unternehmen und Institutionen angewendet. Beispielsweise nutzen Apple und das US Census Bureau Differential Privacy, um datenschutzkonforme Analysen zu ermöglichen.[68] Auch Google integriert Differential Privacy in Dienste wie BigQuery, um die Privatsphäre der Nutzer zu schützen.[69]

Das folgende Beispiel in Tabelle 4-13 zeigt die Anonymisierung eines Datensatzes nach Differential Privacy. Ein App-Entwickler möchte analysieren, wie viele Nutzer bestimmte Funktionen einer App nutzen, ohne die Privatsphäre der Nutzer zu gefährden. Er bedient sich der Differential Privacy und fügt Rauschen hinzu und verschleiert die tatsächliche Nutzung.

Tabelle 4-13: Beispiel für Differential Privacy

ID Nutzer	Nutzung der Chatfunktion	Nutzung der Chatfunktion verrauscht
1	Ja	Ja
2	Nein	Ja
3	Ja	Nein
4	Nein	Nein
5	ja	Ja

Quelle: WIK, Eigene Darstellung.

Im Originaldatensatz nutzen drei Nutzer die betreffende Funktion, im verrauschten Datensatz ebenfalls. Der Schutz funktioniert so, dass jede Antwort mit einer kleinen Wahrscheinlichkeit zufällig geändert wird. Gleichzeitig wird durch einen Algorithmus sichergestellt, dass die Gesamtstatistiken mit hoher Wahrscheinlichkeit nahe an der Wahrheit bleiben.

66 Vgl. Dwork, C. (2006).
67 Vgl. Dwork, C.et al. (2006).
68 https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf [Letzter Abruf 16.12.2024].
69 https://cloud.google.com/bigquery/docs/differential-privacy [Letzter Abruf 16.12.2024].

Entsprechend können Angreifer nicht mit Sicherheit sagen, ob ein bestimmter Nutzer die Funktion tatsächlich nutzt, da das Rauschen die Antwort verschleiert. Aggregierte Analysen können durchgeführt werden. So können z.B. Aussagen gemacht werden wie viele Nutzer diese bestimmte App-Funktion hat bzw. wie verbreitet sie ist, ohne die tatsächlichen Antworten preis zu geben.

4.10 Synthetisierung

Anonymisierung kann auch über die Synthetisierung von Daten hergestellt werden. Dabei werden künstliche, gänzlich neue Datensätze auf Basis statistischer Modelle, die die Eigenschaften der Originaldaten widerspiegeln, erstellt. Diese Technik wird bis dato insbesondere in sensiblen Bereichen wie dem Gesundheitswesen und der Sozialforschung geschätzt.[70]

Die synthetischen Daten werden in einem Modellprozess generiert, der Muster und Zusammenhänge der Originaldaten analysiert und auf Grundlage dieser Erkenntnisse einen strukturähnlichen, aber nicht identischen Datensatz erstellt. Dieser Prozess bewahrt also statistische Merkmale wie Mittelwerte, Standardabweichungen oder Korrelationen, wodurch synthetische Daten für analytische Anwendungen geeignet bleiben. Dies wird insbesondere bei der Entwicklung, dem Testen und der Analyse von Anwendungen und Systemen relevant, bei denen Datenschutz und Datensicherheit von zentraler Bedeutung sind.[71]

Tabelle 4-14: Beispiel für eine Synthetisierung von Daten

Originaldatensatz			
ID Nutzer	**Alter**	**Produkt**	**Kaufhäufigkeit (pro Monat)**
1	35	Buch	2
2	28	Laptop	1
3	42	Haushaltswaren	3
Synthetisierte Daten			
ID Nutzer	**Alter**	**Produkt**	**Kaufhäufigkeit (pro Monat)**
S1	33	Laptop	1
S2	30	Buch	2
S3	40	Haushaltswaren	4

Quelle: WIK, Eigene Darstellung, KI unterstützt.

Tabelle 4-14 zeigt ein Beispiel, das die Anwendung synthetischer Daten bei der Auswertung von Kundendaten zeigt. Aus dem synthetischen Datensatz ist ersichtlich, dass die Eigenschaften des Originaldatensatzes nachgeahmt werden wie etwa das Alter,

70 Vgl. Bindschaedler, V./ Shokri, R. (2016).
71 Vgl. Zur Nutzung von synthetischen Daten im Gesundheitswesen für Studienzwecke. Gonzales, A. et al. (2023).

Produkt und die Kaufhäufigkeit. Die generierten Daten sind dabei vollständig künstlich und es besteht kein Bezug zum Originaldatensatz.

Die Daten sind für analytische Zwecke nutzbar, während der Schutz personenbezogener Informationen sichergestellt ist. Die statistischen Muster von Alter, Präferenz und Kaufhäufigkeit bleiben erhalten. So ist z.B. erkennbar, dass ältere Nutzer zu einer höheren Kaufhäufigkeit tendieren als jüngere Nutzer. Synthetische Daten ermöglichen die Einhaltung der DSGVO[72] und anderer Datenschutzvorschriften, da keine tatsächlichen personenbezogenen Daten verarbeitet werden. Ebenso wird das Risiko von Rückschlüssen oder Re-Identifizierung eliminiert. Im Vergleich zu herkömmlichen Anonymisierungsverfahren bieten synthetische Daten eine höhere Skalierbarkeit. Sie können in beliebigen Mengen und für verschiedene Szenarien generiert werden.

Eine der Herausforderungen der Datensynthetisierung besteht darin, dass synthetische Daten möglicherweise nicht alle relevanten Muster und seltenen, aber wichtigen Ausreißer erfassen, die in den Originaldaten vorhanden sind. Dies kann zu Verzerrungen in den Analyseergebnissen führen, insbesondere wenn es um hochspezifische Anwendungen geht. Die Synthetisierung wird daher oft mit anderen Anonymisierungstechniken wie der Rauschzugabe kombiniert, um die Validität der Analysen zu erhöhen, während die Privatsphäre geschützt bleibt.[73]

4.11 Model-based Obfuscation Knowledge (MOK)

Model-Based Obfuscation Knowledge (MOK) bezeichnet eine innovative Technik zur Anonymisierung sensibler Daten, die gezielt für maschinelles Lernen und statistische Modellierungen entwickelt wurde. Das Konzept basiert darauf, sensible Informationen innerhalb von Daten oder Modellen durch eine bewusste Verzerrung oder Modifikation zu anonymisieren, ohne dabei die Verlässlichkeit und Aussagekraft der Modelle wesentlich zu beeinträchtigen. Die Daten werden also modellbasiert verschleiert. MOK-Daten sind besonders nützlich in Szenarien, in denen die zugrundeliegenden personenbezogenen Informationen durch Modelle verarbeitet werden sollen.

Besonders hervorgehoben wird die Eignung von MOK-Daten für Anwendungen im Bereich des maschinellen Lernens. Sie wurden speziell entwickelt, um realistische, aber anonymisierte Daten bereitzustellen, die sich nahtlos in KI-Trainingsprozesse einfügen lassen, ohne dabei Datenschutzvorgaben zu verletzen. Zudem minimiert die modellbasierte Herangehensweise das Risiko einer Re-Identifikation erheblich.

---

72 Vgl. Europäisches Parlament (2016).
73 Vgl. Charest, A.-S. et al. (2012).

Tabelle 4-15: Model-Based Obfuscation Knowledge (MOK)

Originaldatensatz					
ID Patient	**Alter**	**Geschlecht**	**Diagnose**	**Gewicht (kg)**	**Größe (cm)**
1	35	Männlich	Bluthochdruck	80	175
2	47	Weiblich	Diabetes Typ II	72	162
3	29	Männlich	Asthma	90	180
MOK Datensatz					
ID Patient	**Alter**	**Geschlecht**	**Diagnose**	**Gewicht (kg)**	**Größe (cm)**
1	30–40	Männlich	Hypertonie	82	174
2	40–50	Weiblich	Prädiabetes	74	160
3	20–30	Männlich	Atemwegserkrankung	88	181

Quelle: WIK, Eigene Darstellung, KI unterstützt.

Das folgende Beispiel in Tabelle 4 15 illustriert die Methode anhand von Patientendaten. Modellgestützt wurden exakte Altersangaben durch Kategorien ersetzt, Diagnosen wurden leicht verallgemeinert, ohne den medizinischen Kontext zu verlieren und ein leichtes Rauschen bei Gewicht und Größe hinzugefügt. Dabei wird trotz kleiner Abweichungen die Originalverteilung erhalten, eine Re-Identifikation jedoch erschwert. Im anonymisierten Datensatz wurden die Merkmale ID Patient und Geschlecht unverändert beibehalten, was bedeutet, dass sie nicht anonymisiert wurden. Eine Pseudonymisierung der IDs und zusätzlich eine Generalisierung des Geschlechts (z.B. Gruppe A, Gruppe B, ...) könnten zu einer noch höheren Anonymisierung beitragen.

Dieser Modell-basierte Ansatz stellt sicher, dass die Daten für maschinelles Lernen oder Analyseprozesse nützlich bleiben, ohne dass die Identität der einzelnen Patienten preisgegeben wird.

5 Auswahl geeigneter Verfahren

Literatur und Praxis kennen verschiedene Methoden zur Anonymisierung von Daten, die zuvor dargestellt worden sind. Sie sind einander nicht gleichwertig. Bei der Auswahl von Anonymisierungsverfahren erscheint es daher sinnvoll, Kriterien heranzuziehen, die die Bedürfnisse des jeweiligen Anwendungsfalls spiegeln. Im Folgenden werden einige Kriterien zur Auswahl vorgestellt und diskutiert. Während bei der Vorstellung der Verfahren Anwendungsfälle aus verschiedenen Bereichen herangezogen worden sind, soll hier speziell auf Daten fokussiert werden, die von großen Onlineplattformen (Gatekeepern) bereitgestellt werden. Im Anschluss erfolgen vergleichende Analysen der verschiedenen Verfahren.

5.1 Auswahlkriterien

Gemäß den Vorgaben des DMA[74] sind Gatekeeper dazu verpflichtet, bestimmten Wettbewerbern Daten zugänglich zu machen, um eine faire Wettbewerbsumgebung sicherzustellen. Diese Datenweitergabe muss jedoch so erfolgen, dass die Privatsphäre der betroffenen Nutzer gewahrt bleibt. Damit entsteht ein Spannungsfeld zwischen Datenschutz und Datenverwertbarkeit: Einerseits soll durch geeignete Anonymisierungsverfahren verhindert werden, dass ein Rückschluss auf einzelne Personen möglich ist, andererseits dürfen die so aufbereiteten Daten ihren Nutzen für die analysierenden Wettbewerber nicht vollständig verlieren.

Die Robustheit der Anonymisierung ist dabei entscheidend, um Rückschlüsse auf Individuen auszuschließen. Gleichzeitig darf die Anonymisierung nicht so stark sein, dass die Daten für den ursprünglich beabsichtigten, wettbewerbsfördernden Zweck unbrauchbar werden. Effiziente und nachvollziehbare Verfahren sind gefragt, um mit vertretbarem Aufwand ein hohes Schutzniveau sicherzustellen und regulatorische sowie ökonomische Anforderungen zu erfüllen. Dabei spielen auch die in Kapitel 4 vorgestellten Techniken eine Rolle, da sie im Hinblick auf ihre Eignung bewertet werden müssen. Die in Kapitel 2.3 thematisierte Spannung zwischen Datenschutz und Nutzbarkeit bildet den konzeptionellen Rahmen für diese Bewertung.

Datenschutz/Privacy: Das gewählte Anonymisierungsverfahren muss die Privatsphäre von Individuen wirksam schützen, indem es Rückschlüsse auf personenbezogene Informationen verhindert und das Risiko einer Re-Identifizierung minimiert, auch bei einer Kombination mit anderen Datenquellen.[75] Der Grad der Anonymisierung variiert je nach erforderlichem Sicherheitsniveau und betriebenem Aufwand.

Gatekeeper-Unternehmen unterliegen im Rahmen des DMA strengeren Anforderungen an Anonymisierung und Datenschutz als andere nicht als Gatekeeper benannte

74 Europäisches Parlament (2022b).
75 Vgl. dazu auch die Ausführungen in Kapitel 2.4.

Unternehmen. Insbesondere bei der Weitergabe von Daten an Wettbewerber müssen sie sicherstellen, dass ihre Verfahren regelmäßig evaluiert und an den aktuellen Stand der Datenschutztechnologie angepasst werden. Diese Überprüfungen sind für Gatekeeper verpflichtend und erfolgen in höherer Präzision als bei kleineren Marktteilnehmern.[76]

Rechtliche Konformität: Das Anonymisierungsverfahren muss die rechtlichen Anforderungen der DSGVO[77], des DMA[78] und des DSA[79] vollständig erfüllen. Eine lückenlose Dokumentation der angewandten Anonymisierungsmethoden ist erforderlich, um die Nachvollziehbarkeit des Datenverarbeitungsprozesses sicherzustellen.

Gatekeeper-Unternehmen sind zusätzlich verpflichtet, detaillierte Nachweise über die Wirksamkeit der verwendeten Anonymisierungstechniken zu führen. Im Falle einer Überprüfung müssen sie belegen können, dass die Methoden den gesetzlichen Vorgaben entsprechen und die Anonymisierung so umgesetzt wurde, dass ein Missbrauch der Daten durch Dritte ausgeschlossen ist. Für Wettbewerber, die anonymisierte Daten erhalten, gelten diese Nachweispflichten in geringerem Umfang.

Unverzichtbar ist, dass das gewählte Verfahren in der Lage ist, die rechtlichen Anforderungen zu erfüllen. Es muss den spezifischen Anforderungen sowohl der DSGVO als auch DMA[80] und DSA[81] entsprechen. Drüber hinaus ist das Verfahren der Anonymisierung zu dokumentieren, so dass nachvollziehbar ist, was mit den Daten gemacht worden ist.[82]

Datenintegrität: Für eine aussagekräftige Datenanalyse ist es entscheidend, dass Anonymisierungsverfahren die wesentlichen Merkmale und Korrelationen zwischen den Attributen der Datensätze bewahren. Eine Verletzung der Datenintegrität – das heißt der Richtigkeit, Vertrauenswürdigkeit und Konsistenz von Daten über deren gesamten Lebenszyklus – kann zu unzuverlässigen Analysen, fehlerhaften Entscheidungen und potenziell gravierenden Sicherheitsrisiken führen. Datenintegrität stellt sicher, dass Informationen unverfälscht (frei von Manipulation), kohärent (logisch zusammenhängend) und stabil bleiben, während sie gleichzeitig durch Zugriffs- und Schutzkontrollen vor unbefugten Änderungen oder Verlust gesichert sind. Zusätzlich gewährleistet die Nachvollziehbarkeit, dass sämtliche Modifikationen dokumentiert und überprüfbar sind.[83]

---

76 In der DSGVO werden in Artikel 5 und Artikel 25 Anforderungen an Datenschutz durch, die personenbezogene Daten verarbeiten, einschließlich Gatekeeper. Vgl. Technikgestaltung (Privacy by Design) und die Evaluierung formuliert, was Unternehmen dazu anhält, geeignete technische und organisatorische Maßnahmen mit Rechenschaftsplicht umzusetzen. Diese Vorgaben gelten für alle Unternehmen Europäisches Parlament (2016)

77 Europäisches Parlament (2016).

78 Europäisches Parlament (2022b).

79 Europäisches Parlament (2022c).

80 Europäisches Parlament (2022b).

81 Europäisches Parlament (2022c).

82 Dieses ergibt sich aus der Rechenschaftspflicht in der DSGVO nach Artikel 5(2) du Artikel 24. Vgl. Europäisches Parlament (2016).

83 Vgl. Batini, C./ Scannapieco, M. (2016).

Parallel dazu ist die Datenqualität von zentraler Bedeutung, um sicherzustellen, dass die Daten den Anforderungen eines spezifischen Anwendungsfalls gerecht werden. Hohe Datenqualität zeichnet sich aus durch: Genauigkeit, d.h. korrekte und fehlerfreie Daten, Vollständigkeit, d.h. alle benötigten Datenpunkte sind vorhanden, Konsistenz, d.h. die Daten sind einheitlich über verschiedene Systeme und Formate hinweg, Aktualität der Daten und Relevanz, d.h. die Daten sind für den spezifischen Anwendungsfall geeignet.[84]

Kosteneffizienz und Skalierbarkeit Der technische Aufwand und die Implementierungskosten von Anonymisierungsverfahren sind besonders relevant, wenn große Datenmengen, wie sie bei Gatekeepern vorkommen, verarbeitet werden. Daher ist es entscheidend, dass diese Verfahren einfach skalierbar und möglichst automatisierbar sind, um eine zeitnahe Bereitstellung anonymisierter Daten zu ermöglichen. Auch bei variierenden Datenmengen ist Skalierbarkeit ein wesentlicher Faktor. Darüber hinaus sollten Anonymisierungsverfahren in bestehende Datenmanagement- und Analysetools integrierbar sein, um eine nahtlose Anwendung zu gewährleisten. Schließlich müssen die Kosten der Implementierung und des laufenden Betriebs im Verhältnis zu den Ressourcen und dem Budget des Unternehmens stehen.

Neben den bereits genannten Bewertungskriterien sind auch der **Datentyp und die Sensibilität der Daten** von entscheidender Bedeutung bei der Auswahl eines geeigneten Anonymisierungsverfahrens. Unterschiedliche Datentypen wie Textdaten, Standortdaten, Finanzdaten und Gesundheitsdaten erfordern jeweils spezifische Anonymisierungsstrategien, da sie unterschiedliche Risiken für den Datenschutz darstellen. In diesem Zusammenhang gilt, dass je sensibler die Daten sind, desto höher der erforderliche Anonymisierungsgrad sein sollte. Eine detaillierte Betrachtung der verschiedenen Arten von Daten, die im Kontext der Gatekeeper-Services anfallen, wurde bereits in Kapitel 3.3 durchgeführt. Ebenso wurde auf die verschiedenen Stufen der Sensibilität von Daten eingegangen.

Die Eignung eines Anonymisierungsverfahrens hängt maßgeblich vom spezifischen **Nutzungsszenario** ab, da der analytische Mehrwert der Daten in direktem Zusammenhang mit der Art der Datenanalyse steht. Wie bereits zuvor erläutert, lassen sich dabei zwei zentrale Analyseebenen unterscheiden: die Gruppenebene und die Einzelebene.

Auf der Gruppenebene werden aggregierte Daten betrachtet, um Muster und Trends innerhalb einer definierten Nutzergruppe zu identifizieren. Ein typisches Beispiel ist die Segmentierung von Kundendaten zur Erkennung potenzieller Neukunden oder zur Ableitung allgemeiner Konsumtrends. Der Fokus liegt hierbei nicht auf individuellen Datenpunkten, sondern auf der Gesamtheit der Daten, wodurch die Anforderungen an den Grad der Anonymisierung in der Regel niedriger sind. Demgegenüber erfordert die Einzelebene die Analyse von individuellen Nutzerdaten, etwa zur Direktansprache durch

84 Vgl. Batini, C. et al. (2009).

personalisierte Angebote oder für präzise Vorhersagemodelle. In diesen Fällen müssen die Anonymisierungsverfahren so gewählt werden, dass sie einerseits den Datenschutz gewährleisten und andererseits die notwendigen Detailinformationen für eine zielgerichtete Analyse erhalten bleiben. Dies stellt höhere Anforderungen an die Balance zwischen Datenschutz und Datennutzbarkeit.

Die Stärke der Anonymisierung kann auch je nach **Nutzergruppe** variieren. Abhängig vom Empfänger der anonymisierten Daten, etwa Forschern, Geschäftspartnern oder externen Beratern, können unterschiedliche Anonymisierungsstufen erforderlich sein, um den jeweils benötigten Datenschutz zu gewährleisten.

5.2 Vergleichende Analysen

Gegenstand der folgenden Ausführungen ist die komparative Darstellung der verschiedenen Verfahren aus Kapitel 4 anhand der genannten Kriterien aus 5.1.

Zunächst werden die einzelnen Verfahren anhand der Kriterien Datenschutz, Dateninteгrität, rechtliche Konformität, Kosteneffizienz und Skalierbarkeit sowie Datenqualität und Nutzbarkeit einander tabellarisch gegenübergestellt und die Kriterien über die Verfahren hinweg diskutiert.

Im Anschluss stehen die verschiedenen Arten von Daten im Vordergrund. Die Eignung einzelner Verfahren für spezifische Datenarten wird betrachtet.

Abschließend steht dann die Untersuchung an, welche Art von Analyse auf der Basis, der nach einem spezifischen Verfahren anonymisierten Daten, möglich ist. Hier steht die Prüffrage im Vordergrund, ob der Nachfrager nach Daten diese zum gleichen Zweck verwenden kann, wie der Dateninhaber und so zu den gleichen geschäftsrelevanten Erkenntnissen gelangen kann. Es wird unterschieden zwischen möglichen Analysen auf Einzelebene und auf Gruppenebene.

5.2.1 Beurteilung verschiedener Verfahren anhand von Kriterien

Tabelle 5-1 zeigt die kriterienbasierte Aufbereitung der vorgestellten Verfahren zur Anonymisierung. Die Verfahren werden in Hinblick auf Datenschutz (Privacy), Datenintegrität, rechtliche Konformität, Kosteneffizienz/Skalierbarkeit sowie Datenqualität/Nutzbarkeit bewertet.

Tabelle 5-1: Kriterienbasierte Gegenüberstellung der verschiedenen Ansätze

	Inhalt	Daten-schutz	Datenintegrität	Rechtli-che Konfor-mität	Kosteneffizienz & Skalierbarkeit	Datenqualität/Nutzbarkeit
Suppression	Vollständige Entfernung von Daten	Sehr hoch	Niedrig, da wichtige Daten voll-ständig entfernt werden	Konform	Kosteneffizient, leicht skalierbar	Stark eingeschränkt, ggf. Analysen auf Gruppen-ebene
Maskierung	Abdecken sensibler Daten mit Platz-haltern	Hoch	Mittel bis hoch, jedoch verfälscht	Konform	Kosteneffizient, leicht skalierbar	Eingeschränkt: teilweise Analysen auf Einzel-ebenebene, auf Gruppenebene gut geeignet
Aggregation	Gruppierung von Daten mit weniger Details	Mittel	Mittel/eingeschränkt, Datenstruktur bleibt nicht erhalten	konform	Sehr kosteneffizient, gut skalierbar	Eingeschränkt, keine Analysen auf Einzelebene, nur auf Gruppenebene
Noise Addition	Hinzufügen von Zufallsrauschen	Mittel bis hoch	Mittel bis hoch, abhängig von der Dosierung des Rauschens	Konform	Relativ kosteneffizient, gut skalierbar	Eingeschränkt, eingeschränkte Analysen auf Ein-zelebene, auf Gruppenebene Analysen möglich
Randomi-sierung	Zufällige Veränderung von Daten	Mittel bis hoch	Mittel bis niedrig, Datenstruktur bleibt erhalten	Konform	Relativ kosteneffizient, gut skalierbar	Eingeschränkt, keine Analysen auf Einzelebene, Gruppenebene Analysen möglich
Permutation	Vertauschen von Werten einzelner At-tribute zwischen Datensätzen	Mittel	Niedrig, deutlich eingeschränkt, Beziehungen gehen verloren	Konform	Kosteneffizient, skalierbar	Begrenzt, keine Analysen Einzelebene, nur auf Gruppenebene
Swapping	Austausch sensibler Attribute über Datensätze hinweg	Mittel	Mittel, Verzerrungen, bleibt be-grenzt erhalten, für einige Analy-sen problematisch	Konform	Kosteneffizient, gut ska-lierbar	Variiert, je nach Anwendungsfall keine Analysen auf Einzelebene, nur auf Gruppenebene
Hashing	Verschlüsselung in eindeutige, nicht rekonstruierbare Werte (z.B. SH-256)	Sehr hoch	Hoch, der gleiche Wert führt immer wieder zum gleichen Hash-Wert	Konform	Kostengünstig, gut ska-lierbar	Sehr eingeschränkt, nur exakte Vergleiche mög-lich, für analytische Zwecke unbrauchbar
K-Anonymität	Gruppierung von Daten, so dass Per-son in Menge von k nicht unter-scheidbar	Mittel bis hoch	Mittel, da Generalisierung Details reduziert	Konform	Hohe Effizienz bei gro-ßen Datensätzen.	Eingeschränkte Analysen auf Einzelebene, gut geeignet für Analysen auf Gruppenebene, Ab-hängig von Höhe k Informationsverlust
L-Diversität	Gruppe mit identischen Merkmalen hat mindestens l verschiedene, sen-sible Werte	Sehr hoch	Mittel bis niedrig, mehr Generali-sierung, was Detailverlust erhöht, Verzerrung	Konform	Kostengünstig, je nach Datensatzgröße skalier-bar.	Deutlich eingeschränkte Analysen auf Einzel-ebene, gut geeignet für Analysen auf Gruppen-ebene, abhängig von Größe des Datensatzes
T-Closeness	Verteilung des sensiblen Attributs in-nerhalb jeder Gruppe weicht um fest-gelegtes T von Gesamtverteilung ab	Sehr hoch	Niedrig, noch stärkere Generalisie-rung, starke Verzerrungen	Konform	Skalierbarkeit begrenzt, je nach Datensensibilität.	Stark eingeschränkte Analysen auf Einzelebene, sehr gut geeignet für Analysen auf Gruppen-ebene durch realistischere Verteilung
Differential Privacy	Hinzufügen von Zufallsrauschen	Sehr hoch	Mittel bis niedrig, Verfälschung durch Rauschen, Trends bleiben	Konform	Hängt von Parametern ab, skalierbar.	Eingeschränkt bei Analysen auf Einzelebene, gut geeignet für Analysen auf Gruppenebene
Synthetische Daten	Vollständig künstliche Daten mit ähn-lichen Mustern wie reale Daten	Sehr hoch	Abhängig von Modell, Muster blei-ben erhalten, Verzerrungen mög-lich	Konform	Kosten hoch für die Ge-nerierung, gut skalierbar	Abhängig vom Modell, nicht realitätsentspre-chend, Analysen auf Einzel- und Gruppenebene
MOK Daten	Modellbasierte Verschleierung	Sehr hoch	Sehr hoch bei gut trainierten Mo-dellen	Konform	Variabel, oft kostenin-ten-siv, gut skalierbar	Eingeschränkt für Analysen auf Einzelebene, sehr gut für Analysen auf Gruppenebene

Quelle: WIK, Eigene Analyse.

Verfahren wie Suppression und Hashing bieten im Regelfall ein sehr hohes **Datenschutzniveau**, da sie direkte Identifizierungsmerkmale entfernen oder unwiderruflich transformieren. Allerdings leidet darunter oft die Nutzbarkeit der Daten. Auch klassische Methoden wie K-Anonymität, L-Diversität und T-Closeness sowie Differential Privacy erreichen einen sehr hohen Datenschutzstandard, indem sie sicherstellen, dass ein Rückschluss auf individuelle Personen selbst unter Kombination mit weiteren Datenquellen erschwert wird. Synthetische Daten und modellbasierte Verfahren (MOK-Daten) sind ebenfalls sehr robust in Bezug auf Privatsphäre, da sie nur Muster der Originaldaten abbilden und somit keine direkten Re-Identifikationen zulassen.

Bei Suppression, Permutation oder reiner Aggregation können wesentliche Zusammenhänge und Korrelationen in den Daten verloren gehen, was die **Datenintegrität** im Sinne der Erhaltung von Strukturen und Beziehungen stark beeinträchtigt. Noise Addition und Randomisierung verändern zwar einzelne Werte, erhalten aber oft grundlegende Zusammenhänge, sofern das Rauschen angemessen dosiert ist. K-Anonymität, L-Diversität und T-Closeness sowie Differential Privacy ermöglichen eine kontrollierte Verzerrung, bei der Strukturen zu einem gewissen Grad erhalten bleiben. Modellbasierte Ansätze (MOK-Daten) und synthetische Daten versuchen, die internen Zusammenhänge weitgehend beizubehalten, indem sie neue Daten generieren, die ähnliche statistische Muster wie die Originaldaten zeigen.

Die meisten der genannten Verfahren verfügen über eine **rechtliche Konformität** mit DSGVO, DMA und DSA, sofern sie korrekt angewendet und dokumentiert werden. Verfahren wie Suppression, Hashing oder Aggregation sind konzeptionell einfach und daher leicht nachvollziehbar. Die Nachvollziehbarkeit ist wichtig für Gatekeeper, die detaillierte Nachweise erbringen müssen. Verfahren mit höherer Komplexität, wie Differential Privacy oder synthetische Datenerzeugung, erfüllen zwar die rechtlichen Anforderungen, setzen aber eine exakte Dokumentation und möglicherweise externe Audits voraus, um ihre Wirksamkeit zu belegen. Die regulatorische Konformität hängt hier vor allem von der korrekten Parametrisierung (z. B. Wahl von ε bei Differential Privacy) und der lückenlosen Dokumentation ab.

Eher einfache Verfahren wie Suppression, Maskierung, Aggregation oder Hashing haben in der Regel eine hohe **Kosteneffizienz** und **Skalierbarkeit**. Sie lassen sich leicht automatisieren und in bestehende Systeme integrieren. Aufwändigere Methoden wie synthetische Datenerzeugung oder komplexe Modellierungen (MOK-Daten) sind im Regelfall teurer, sowohl in der Entwicklung als auch im laufenden Betrieb. Auch Differential Privacy kann je nach Komplexität der Implementierung höhere Kosten verursachen, ist aber dennoch in großen Datensätzen relativ gut skalierbar, wenn die entsprechenden Infrastrukturen vorhanden sind. Die Kosten-Nutzen-Abwägung wird für Gatekeeper daher wichtig, um sicherzustellen, dass die Investitionen in komplexere Verfahren einen Mehrwert für die Wettbewerbsfähigkeit und die Rechtskonformität bieten.

Bei Verfahren wie Suppression oder starker Aggregation sinken die **Datenqualität** und **Nutzbarkeit** signifikant, da wesentliche Details für wettbewerbsrelevante Analysen verloren gehen. Maskierung, Noise Addition oder Randomisierung erhalten zwar die Möglichkeit, gewisse statistische Aussagen abzuleiten, beeinträchtigen aber dennoch die Genauigkeit. Verfahren aus der Familie der K-Anonymität (L-Diversität, T-Closeness) oder Differential Privacy versuchen ein Gleichgewicht zu finden: Sie reduzieren das Re-Identifikationsrisiko, ohne die grundlegenden Muster der Daten vollständig zu zerstören. Synthetische Daten und MOK-Daten haben hier ein großes Potenzial, da sie unter optimalen Bedingungen komplexe Zusammenhänge nachbilden und eine hohe Nutzbarkeit ermöglichen, allerdings meist zu höheren Kosten und mit größerem technischem Aufwand.

Auch bei der Zuordnung der Verfahren zu den Kriterien zeigt sich das Spannungsfeld aus Datenschutz, Datenqualität und Kosten. Einfache Verfahren wie Suppression oder Hashing bieten hohen Datenschutz, sind rechtlich konform und skalierbar, aber beeinträchtigen stark die Datenqualität. Komplexere Verfahren wie Differential Privacy, K-Anonymität, L-Diversität, T-Closeness oder synthetische Datenerzeugung bieten ein besseres Gleichgewicht zwischen Privatsphäre und Nutzbarkeit, sind aber in der Implementierung aufwändiger und potenziell kostspieliger.

Für Gatekeeper ist die Wahl des Verfahrens vor dem Hintergrund der Vorgaben des DMA zentral: Sie müssen sowohl die rechtliche Konformität als auch hohe Datenintegrität und Nutzbarkeit sicherstellen. Während einfache Methoden den regulatorischen Pflichten rasch nachkommen, können sie die Datenqualität so stark mindern, dass der Wettbewerbszweck untergraben wird. Aus Sicht der Gatekeeper hingegen könnten diese Verfahren daher auch aus strategischen Gründen gewählt werden. Sie kommen so den Datenschutzverpflichtungen nach, haben niedrige Kosten und stellen Daten zur Verfügung, auf deren Basis nur eingeschränkt Analysen möglich sind. Höherentwickelte Verfahren wie Differential Privacy oder synthetische Daten ermöglichen eine bessere Balance, erfordern aber eine sorgfältige Parametrisierung, höhere Investitionen und ein fundiertes technisches Know-how. Somit hängt die optimale Entscheidung stark von den spezifischen Anforderungen, der Größe und Art der Datenbestände sowie den technischen und rechtlichen Ressourcen des Gatekeepers ab.

5.2.2 Eignung verschiedener Verfahren nach Datenart

Tabelle 5-2 bietet einen systematischen Überblick über gängige Anonymisierungstechniken, die je nach Art der zu schützenden Daten angewendet werden können. Dabei zeigt sich, dass die Auswahl der Methode maßgeblich von den datenspezifischen Eigenschaften abhängt.

Persönliche Identifikationsdaten wie Namen oder E-Mail-Adressen werden häufig durch Suppression (vollständige Entfernung) oder Hashing geschützt, da diese Verfahren eine direkte Rückverfolgung erschweren und dennoch eine gewisse Datenintegrität

erhalten. Dies ist für die Wahrung der Privatsphäre essenziell und wird in der Praxis, etwa bei großen Datensätzen sehr häufig angewandt.[85]

Bei Informationen, die nicht unmittelbar zur Personenidentifikation führen, wie **Benutzernamen** oder **Profilbilder**, kommen Maskierungstechniken zum Einsatz. Hierbei werden konkrete Werte durch Platzhalter ersetzt, um die Datennutzbarkeit hinsichtlich Struktur und Grundfunktionen zu erhalten, ohne jedoch die Identität offenzulegen. kommen sensiblere Informationen wie **Zahlungsdaten** ins Spiel, empfiehlt sich ein Hashing oder Swapping. Während Hashing eine sichere Einwegfunktion bereitstellt, durch die eine Rückrekonstruktion so gut wie unmöglich ist, nimmt Swapping einen Austausch sensibler Attribute zwischen Datensätzen vor.

Bei **Geolokalisierungsdaten** oder IP-Adressen wird häufig auf Randomisierung und T-Closeness zurückgegriffen. Randomisierung durch Hinzufügen von zufälligen Abweichungen erschwert die Re-Identifikation. T-Closeness stellt sicher, dass die Verteilung sensibler Werte in jeder Anonymitätsgruppe dem gesamten Datensatz möglichst ähnlich bleibt. Dadurch lässt sich das Risiko verringern, dass Einzelpersonen durch spezifische Ausprägungen bestimmter Merkmale wiedererkannt werden. Bei **Geräteinformationen** wie Betriebssystem oder Browsertyp ist T-Closeness ebenfalls geeignet, da dadurch wichtige Muster erhalten bleiben, jedoch eine individuelle Zuordnung auf ein bestimmtes Gerät mit hoher Wahrscheinlichkeit verhindert wird.

Interaktions- und Verhaltensdaten, etwa Klickverhalten oder Suchhistorien, unterliegen besonderen Herausforderungen, weil hier komplexe Muster und Korrelationen eine Rolle spielen. Noise Addition (Zufallsrauschen) und Differential Privacy gelten als Standard. Durch die Hinzufügung kontrollierter statistischer Störungen (Noise) lässt sich ein formal quantifizierbares Schutzniveau erzielen. Differential Privacy setzt dazu meist auf ein Epsilon (ε), das die Privatsphärenstärke angibt. MOK-Daten (modellorientierte Kodierung) ermöglichen es zudem, komplexe Verhaltensmuster in synthetische oder abstrahierte Strukturen zu übersetzen, ohne direkt personenbezogene Informationen zu offenbaren.

Bei **Kommunikationsmetadaten** wie Zeitstempeln oder Netzwerkverbindungen kann Permutation helfen. Hier werden Werte untereinander vertauscht, um eindeutige Zuordnungen zu unterbinden. Im Kontext von **Kauf- und Transaktionsdaten** kommen vermehrt Swapping und K-Anonymität zum Einsatz. K-Anonymität sorgt dabei dafür, dass jede Person innerhalb einer Gruppe von mindestens k unterscheidbar ähnlichen Datensätzen auftritt, was Re-Identifikationsrisiken drastisch senkt. Üblicherweise werden k-Werte von 3 bis 10 angewandt, um ein ausreichendes Anonymitätsniveau zu gewährleisten.

---

85 Vgl. Sweeney (2002b); Barker, E./ Harris, J. (2012).

Bei besonders sensiblen und dynamischen Datentypen wie **Sensor-, Kontext- oder Netzwerkdaten** und **Werbeinteraktionsdaten** ist eine stärkere Verfälschung oder Gruppierung angebracht. L-Diversität, K-Anonymität und MOK-Daten sorgen hier dafür, dass weder einzelne Knoten in Netzwerken noch spezifische Präferenzen bei Werbeanzeigen rückführbar auf einzelne Individuen sind. Gleichzeitig bleibt das Muster der Daten für Analysezwecke weitgehend intakt.

Tabelle 5-2: Eignung verschiedener Verfahren nach Datenart

Datenart	Geeignete Anonymisierungstechniken	Erläuterung
Persönliche Identifikationsdaten (Name, E-Mail, Telefonnummer)	Suppression, Hashing	Vollständige Entfernung schützt die Privatsphäre; Hashing verhindert die Rekonstruktion, bewahrt aber Datenintegrität.
Benutzernamen und Profilbilder	Maskierung	Platzhalter ersetzen sensible Daten, was die Struktur erhält, und grundlegende Nutzbarkeit ermöglicht.
Zahlungsinformationen (Kreditkarten-, Bankdaten)	Hashing, Swapping	Hashing schützt vor Rekonstruktion, Swapping tauscht sensible Attribute aus, um Datenschutz zu stärken.
Geolokalisierungsdaten (GPS-Koordinaten, IP-Adressen)	Randomisierung, T-Closeness	Zufällige Veränderung erschwert Rückverfolgung; T-Closeness wahrt Verteilungsmuster und reduziert Re-Identifikationsrisiken.
Geräteinformationen (Betriebssystem, Modell, Browsertyp)	T-Closeness	Bewahrt statistische Muster und minimiert das Risiko, einzelne Geräte zu identifizieren.
Interaktions- und Verhaltensdaten (Klickverhalten, Suchhistorien)	Noise Addition, Differential Privacy, MOK Daten	Noise Addition schützt durch Zufallsrauschen, Differential Privacy bietet Schutz bei aggregierten Analysen. MOK-Daten bewahren Muster und Korrelationen durch modellbasierte Transformation.
Kommunikationsdaten (Zeitstempel, Metadaten von Nachrichten)	Permutation	Vertauschen von Werten erschwert die Rückverfolgbarkeit auf Einzelpersonen.
Kauf- und Transaktionsdaten (Kaufverläufe, Warenkorbdaten)	Swapping, K-Anonymität	Swapping schützt sensible Details; K-Anonymität bildet Gruppen ähnlicher Muster, die nicht voneinander unterscheidbar sind.
Sensor- und Kontextdaten (Bewegungs-, Aktivitätsdaten)	Synthetische Daten, MOK-Daten	Synthetische Daten generieren realistische, aber künstliche Muster, ohne echte personenbezogene Daten zu nutzen. MOK-Daten ermöglichen die Anonymisierung sensibler Muster uns sind gut skalierbar, auch für große Datenmengen.
Verbindungs- und Netzwerkdaten (Freunde, Gruppen, Interaktionen)	L-Diversität, MOK-Daten, K-Anonymität	L-Diversität und K-Anonymität stellen sicher, dass sensible Daten innerhalb einer Gruppe vielfältig sind, um Re-Identifikation zu minimieren. MOK-Daten können kausale Beziehungen erhalten.
Werbeinteraktionsdaten (gesehen, angeklickt, gemieden)	Noise Addition, Differential Privacy	Noise Addition schützt durch Zufallsrauschen; Differential Privacy ermöglicht sichere aggregierte Analysen und personalisierte Werbung.

Quelle: WIK, Eigene Analyse.

Insgesamt wird deutlich, dass eine sorgfältige Auswahl der Anonymisierungstechnik wichtig ist, um den spezifischen Anforderungen unterschiedlicher Datentypen gerecht zu werden. Die in der Tabelle genannten Techniken sind etablierte Methoden, deren Wirksamkeit und Effizienz belegt sind. Die Kombination mehrerer Verfahren, wie beispielsweise die Verbindung von Noise Addition mit Differential Privacy oder von Hashing mit Swapping, die Robustheit erhöhen und zu einem nachhaltig hohen Schutzniveau führen.

5.2.3 Eignung verschiedener Verfahren nach Nutzungsszenario

Tabelle 5-3 zeigt, wie sich verschiedene Anonymisierungsverfahren auf die Durchführbarkeit von Analysen auf Einzel- und Gruppenebene auswirken. Während einige Methoden eine Nutzung auf der Einzelebene nahezu unmöglich machen, bieten sie auf Gruppenebene weiterhin ein hohes Maß an Verlässlichkeit. Andere Techniken erhalten zwar die Möglichkeit zu Analysen auf der Einzelebene teilweise, beeinträchtigen jedoch deren Genauigkeit.

Verfahren wie Suppression und Hashing sind auf Einzelebene problematisch, da durch die Entfernung oder Umwandlung der Originalwerte individuelle Analysen kaum mehr möglich sind. Auf Gruppenebene dagegen bleiben die Daten zumindest für grobe Mustererkennungen nutzbar, da sich die fehlenden Informationen durch die Aggregation relativieren. Ähnlich verhält es sich bei Verfahren wie Aggregation, wo bereits auf Dateneingangsebene keine Einzelinformationen mehr vorliegen, wodurch individuelle Analysen ausgeschlossen sind. Für Gruppenanalysen ist diese Technik jedoch geeignet, da sie konsistente statistische Muster bewahrt.

Verfahren wie Maskierung, K-Anonymität oder L-Diversität sorgen dafür, dass individuelle Datensätze zwar in gewisser Weise verändert oder verallgemeinert werden, die Gesamtstruktur des Datensatzes aber weitgehend erhalten bleibt. Dadurch sind Einzelwertanalysen, wenn auch eingeschränkt, noch möglich. Auf Gruppenebene bleiben hingegen umfangreiche analytische Fähigkeiten erhalten: Muster, Trends und Verteilungen lassen sich nach wie vor erkennen.

Noise Addition, Randomisierung, Permutation oder Swapping erschweren präzise Einzelwertanalysen, da durch das Hinzufügen von Rauschen oder die Vertauschung von Werten der ursprüngliche Bezug des Datensatzes gestört wird. Dennoch bleiben Gruppenauswertungen weitgehend stabil, denn die statistischen Eigenschaften verändern sich nur in begrenztem Ausmaß. Auch Differential Privacy, bei der der Schutz der Privatsphäre über gezielt eingesteuertes Rauschen erfolgt, schränkt Einzelanalysen zugunsten stabiler Gruppenstatistiken ein.

Synthetische Daten und MOK-Daten sind wenig realitätsnah. Bei synthetischen Daten werden vollständig künstliche Datensätze generiert, die lediglich statistische Ähnlichkeiten mit dem Original aufweisen. Einzelanalysen sind in diesem Fall nicht sinnvoll, da keine realen Werte mehr existieren. Für Gruppenanalysen hingegen bieten synthetische Daten ein gutes Instrument, um Muster und Korrelationen zu untersuchen, ohne die Privatsphäre realer Personen zu gefährden. MOK-Daten (Modell-Orientierte Kodierung) verfolgen einen ähnlichen Ansatz: Auch hier werden die Daten so transformiert, dass Einzelwertanalysen zwar stark eingeschränkt, Gruppenmuster aber erhalten bleiben.

Tabelle 5-3: Eignung verschiedener Verfahren nach Nutzungsszenario

Anonymisierungsverfahren	Analysen auf Einzelebene	Analysen auf Gruppenebene
Suppression	Nicht möglich: Entfernte Daten machen Analysen auf individueller Ebene unmöglich.	Gut geeignet: Solange die entfernten Daten für Gruppenanalysen irrelevant sind.
Maskierung	Teilweise möglich: Struktur bleibt erhalten, aber Originalwerte fehlen.	Sehr gut geeignet: Maskierung hat keinen Einfluss auf aggregierte Ergebnisse.
Aggregation	Nicht möglich: Einzelne Werte gehen durch Gruppierung verloren.	Sehr gut geeignet: Perfekt für gruppenbasierte Analysen, Muster bleiben erhalten.
Noise Addition	Eingeschränkt möglich: Rauschen beeinträchtigt Genauigkeit einzelner Werte.	Gut geeignet: Statistische Muster bleiben trotz Rauschen erhalten.
Randomisierung	Eingeschränkt möglich: Originalwerte sind durch Randomisierung stark verändert.	Gut geeignet: Gruppenstatistiken bleiben intakt, wenn Randomisierung moderat erfolgt.
Permutation	Eingeschränkt möglich: Reihenfolgeänderungen machen Analysen auf individueller Ebene schwierig.	Gut geeignet: Statistische Eigenschaften werden nicht beeinträchtigt.
Swapping	Eingeschränkt möglich: Werte sind vertauscht, individuelle Präzision leidet.	Gut geeignet: Aggregierte Muster bleiben erhalten, da Verteilung konsistent bleibt.
Hashing	Nicht möglich: Nur exakte Vergleiche von Hashwerten möglich, keine Analysen.	Nicht möglich: Aggregation ist nicht sinnvoll mit Hashwerten.
K-Anonymität	Teilweise möglich: Abhängig von der Generalisierung; Daten sind weniger präzise.	Sehr gut geeignet: Aggregationen sind robust und identifizierende Daten fehlen.
L-Diversität	Teilweise möglich: Bietet zusätzlichen Schutz, präzise Analysen eingeschränkt.	Sehr gut geeignet: Stärkerer Schutz, aber gute Nutzbarkeit für Gruppenanalysen.
T-Closeness	Eingeschränkt möglich: Einschränkungen erhöhen den Schutz, verringern Präzision.	Sehr gut geeignet: Statistische Eigenschaften bleiben innerhalb enger Grenzen erhalten.
Differential Privacy	Eingeschränkt möglich: Privatsphärenschutz führt zu Ungenauigkeiten auf Einzelwertebene.	Gut geeignet: Gruppenstatistiken bleiben durch kontrollierte Rauschstärke intakt.
Synthetische Daten	Nicht möglich: Keine Originaldaten; synthetisierte Werte sind Simulationen.	Gut geeignet: Modelle erhalten aggregierte Muster und statistische Eigenschaften.
MOK-Daten	Eingeschränkt möglich: Modellgenerierte Daten weichen vom Original ab.	Gut geeignet: Modelle können statistische Beziehungen auf Gruppenebene erhalten.

Quelle: WIK, Eigene Analyse.

Insgesamt lässt sich festhalten, dass Techniken, die den Datenschutz auf Einzelebene stärken, meist die Präzision von Einzelanalysen reduzieren. Gleichzeitig sind die meisten Verfahren gut oder sehr gut geeignet, um auf Gruppenebene weiterhin aussagekräftige Analysen zu ermöglichen. Die Wahl der richtigen Anonymisierungsmethode hängt somit maßgeblich von den Zielen der Datenanalyse ab: Geht es um individuelle Auswertungen, müssen entsprechende Verfahren sorgfältig gegen den gewünschten Privatsphärenschutz abgewogen werden. Stehen dagegen aggregierte Kennzahlen und statistische Gesamtmuster im Vordergrund, eignen sich viele Verfahren sehr gut, um gleichzeitig Privatsphäre und Analysequalität zu gewährleisten.

6 Schlussbetrachtung und Ausblick

Im Fokus der Untersuchung stand das komplexe Spannungsfeld aus Datenschutz, der Verwendbarkeit anonymisierter Daten und der Verhältnismäßigkeit verschiedener Anonymisierungsverfahren. Die aufgeworfenen Forschungsfragen konnten umfassend beantwortet werden.

Zunächst wurde deutlich, dass die Interessenlagen von Dateninhabern und Datenempfängern stark divergieren. Dateninhaber bevorzugen einen hohen Grad an Anonymisierung im Gegensatz zu Datenempfängern, die eine geringere Anonymisierung favorisieren, da sie eine hohe Datenqualität und bessere Nutzbarkeit für ihre Analysen wünschen.

Der Schwerpunkt der anschließenden Ausführungen lag auf den verschiedenen Anonymisierungsverfahren, ihren Stärken, Schwächen und der praktischen Umsetzung. Neben einer ausführlichen Darstellung ihrer Funktionsweisen wurden vergleichende Analysen durchgeführt.

Es konnte gezeigt werden, dass die Auswahl des passenden Verfahrens maßgeblich von den Anwendungszielen, den Eigenschaften der verfügbaren Daten und dem technischen sowie ökonomischen Kontext abhängt. Einfache Techniken erlauben die einfache Einhaltung regulatorischer Vorgaben, beeinträchtigen jedoch häufig die Datenqualität. Moderne Verfahren bieten bessere Kompromisse, erfordern aber vertiefte Fachkenntnisse, höhere Investitionen und eine sorgfältige Parametrisierung. Für Gatekeeper stellt sich diese Entscheidung auch in strategischer Hinsicht, um den rechtlichen Anforderungen des DMA nachzukommen, ohne die eigene Wettbewerbsposition zu gefährden.

Aus regulatorischer Sicht sind zwei Fragen, in deren Zusammenhang Kriterien wie die Nachvollziehbarkeit der gewählten Techniken, die Eignung zur Datennutzung durch Dritte und die rechtskonforme Umsetzung an Bedeutung gewinnen, entscheidend:

- Ist die Beschwerde über ein „Zuviel“ an Anonymisierung gerechtfertigt?
- Kann der Datenempfänger die Daten zu dem gleichen Zweck verwenden wie der Dateninhaber und kann er die gleichen geschäftsrelevanten Erkenntnisse erlangen?

Künftige Forschung sollte nicht nur bestehende Verfahren weiterentwickeln, sondern auch neue Modelle und Regulierungsansätze hervorbringen, um den Spannungsbogen zwischen Datenschutz und Datenverwendbarkeit noch besser aufzulösen.

Insgesamt zeigt die Untersuchung, dass ein Zugang zu Daten nur dann sinnvoll umgesetzt werden kann, wenn geeignete Anonymisierungsverfahren mit klaren regulatorischen Strukturen und einer integrierten Betrachtung der Interessen der beteiligten Akteure kombiniert werden. Nur auf diese Weise lässt sich der ökonomische Wert der Daten maximieren, während gleichzeitig die Privatsphäre der Betroffenen gewahrt bleibt.

7 Literaturverzeichnis

Aichroth, A. et al (2020): Anonymisierung und Pseudonymisierung von Daten für Projekte des maschinellen Lernens, Bitkom. Online verfügbar unter https://www.bitkom.org/sites/main/files/2020-10/201002_lf_anonymisierung-und-pseudonymisierung-von-daten.pdf [Letzter Abruf 25.11.2024].

Arnaut, C. et al. (2018): Study on data sharing between companies in Europe. A study prepared for the European Commission. Online verfügbar unter https://publications.europa.eu/resource/cellar/2d6d436e-4832-11e8-be1d-01aa75ed71a1.0002.01/DOC_1 [Letzter Abruf 2.12.2024].

Arnold, R. et al. (2020): European data economy: Between competition and regulation, Bad Honnef. Online verfügbar unter https://www.wik.org/fileadmin/Studien/2021/European_Data_Economy.pdf [Letzter Abruf 12.12.2024].

Barker, E./ Harris, J. (2012): NIST SP 800-107 Revision 1: Recommendation for Applications Using Approved Hash Algorithms, National Institute of Standards and Technology (NIST), Online verfügbar unter https://csrc.nist.gov/publications/detail/sp/800-107/rev-1/final [Letzter Abruf 17.12.2024].

Batini, C. et al. (2009): Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, 3, Online verfügbar unter https://www.researchgate.net/profile/Cinzia-Cappiello/publication/220565749_Methodologies_for_Data_Quality_Assessment_and_Improvement/links/0c9605294d48ba36ac000000/Methodologies-for-Data-Quality-Assessment-and-Improvement.pdf [Letzter Abruf 16.12.2024].

Batini, C./ Scannapieco, M. (2016): Data and Information Quality: Dimensions, Principles and Tech-niques, Berlin, Heidelberg.

Batura O. et al. (2023): The emergence of non-personal data markets. Study requested by the ITRE committee of the European Parliament. PE 740.098. Online verfügbar unter https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740098/IPOL_STU(2023)740098_EN.pdf [Letzter Abruf 12.12.2024].

Bender, A. (2015): Anwendbarkeit von Anonymisierungstechniken im Bereich Big Data (Doctoral dissertation, Master Thesis, KIT, Karlsruhe, Germany). Online verfügbar unter https://www.inovex.de/wp-content/uploads/anwendbarkeit-von-anonymisierungstechniken-im-bereich-big-data-andreas-bender-Mai-2015.pdf [Letzter Abruf 26.11.2024].

Bindschaedler, V./ Shokri, R. (2016): Synthesizing Plausible Privacy-Preserving Location Traces, 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2016, pp. 546-563, Online verfügbar unter https://ieeexplore.ieee.org/document/7546522 [Letzter Abruf 26.11.2024].

BVDW (2018): Datenwertschöpfung und Qualität von Daten, Bundesverband Digitale Wirtschaft (BVDW) e.V., Düsseldorf.

Charest, A.-S. et al. (2012): The use of synthetic data to prevent disclosure, Statistical Journal of the IAOS, 28(3–4), 133–145.

Cichy, C./ Rass, S. (2019): An Overview of Data Quality Frameworks, IEEE Access. Online verfügbar unter https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813 [Letzter Abruf 30.09.2024].

Cornes, R./ Sandler, T. (1986): The Theory of Externalities, Public Goods and Club Goods. Cambridge University Press.

Curry, E. (2016): The Big Data Value Chain: Definition, Concepts, and Theoretical Approaches, in Cavanillas, J., Curry, E., Wahlster, W. (eds) New Horizons for a Data-Driven Economy. Springer, pp. 29-38.

Domingo-Ferrer, J./ Torra, V. (2008): A Critique of k-Anonymity and Some of Its Enhancements. Online verfügbar unter https://www.researchgate.net/publication/220265538_A_Critique_of_k-Anonymity_and_Some_of_Its_Enhancements [Letzter Abruf 28.10.2024].

Duncan, G. T./ Lambert, D. (1989): The Risk of Disclosure for Microdata, in: Journal of Business and Economic Statistics 7(2), p. 207-217.

Dwork, C. (2006): Differential Privacy, in: Bugliesi, M. et al. (Hrsg.): Automata, Languages and Programming, Lecture Notes in Computer Science, vol 4052, Berlin, Heidelberg, p. 1-12.

Dwork, C.et al. (2006): Calibrating noise to sensitivity in private data analysis, Online verfügbar unter https://uvammm.github.io/docs/dwork.pdf [Letzter Abruf 16.12.2024].

El Emam, K. et al. (2011). The re-identification risk of Canadians from longitudinal demographics. BMC Medical Informatics and Decision Making, 11, 46., online verfügbar https://doi.org/10.1186/1472-6947-11-46 [Letzter Abruf 12.12.2024].

Europäische Kommission (2020): Eine europäische Datenstrategie. Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen, COM(2020) 66 final, Online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52020DC0066 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2010): Richtlinie 2010/40/EU des Europäischen Parlaments und des Rates vom 7. Juli 2010 zum Rahmen für die Einführung intelligenter Verkehrssysteme im Straßenverkehr und für deren Schnittstellen zu anderen Verkehrsträgern, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32010L0040 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2015): Richtlinie (EU) 2015/2366 des Europäischen Parlaments und des Rates vom 25. November 2015 über Zahlungsdienste im Binnenmarkt, zur Änderung der Richtlinien 2002/65/EG, 2009/110/EG und 2013/36/EU und der Verordnung (EU) Nr. 1093/2010 sowie zur Aufhebung der Richtlinie 2007/64/EG. Online verfügbar unter https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015L2366.

Europäisches Parlament (2016): Verordnung (EU) 2016/679 des Europäische Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2018): Verordnung (EU) 2018/1807 des Europäische Parlaments und des Rates vom 14. November 2018 über einen Rahmen für den freien Verkehr nicht-personenbezogener Daten in der Europäischen Union, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32018R1807 [Letzter Abruf 12.12.20204].

Europäisches Parlament (2019): Richtlinie 2019/1024 des Europäische Parlaments und des Rates vom 20. Juni 2019 über offene Daten und die Weiterverwendung von Informationen des öffentlichen Sektors, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32019L1024 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2022a): Verordnung (EU) 2022/868 des Europäischen Parlaments und des Rates vom 30. Mai 2022 über europäische Daten-Governace und zur Änderung der

Verordnung (EU) 2018/1724 (Daten-Governance-Rechtsakt), online verfügbar https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32022R0868

Europäisches Parlament (2022b): Verordnung (EU) 2022/1925 des Europäischen Parlaments und des Rates vom 14.September 2022 über bestreitbare und faire Märket im digitalen Sektor und zur Änderung der Richtlinien (EU) 2019/1937 und (EU) 2020/1828, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32022R1925 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2022c): Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32022R2065 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2023): Verordnung (EU) 2023/2854 des Europäischen Parlaments und des Rates vom 13. Dezember 2023 über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung sowie zur Änderung der Verordnung (EU) 2017/2394 und der Richtlinie (EU) 2020/1828, online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202302854&qid=1724841321851 [Letzter Abruf 12.12.2024].

Europäisches Parlament (2024): Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828, online verfügbar unter https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF [Letzter Abruf 12.12.2024].

European Commission (2018), Directorate-General for Communications Networks, Content and Technology, Scaria, E., Berghmans, A., Pont, M., Arnaut, C., et al., Study on data sharing between companies in Europe : Final report, Publications Office, 2018, https://data.europa.eu/doi/10.2759/354943

European Commission (2018), Directorate-General for Communications Networks, Content and Technology, Wauters, P.; Siede, A.; Cocoru, D.; Linz, F., et al.: Study on emerging issues of data ownership, interoperability, (re-)usability and access to data, and liability: Final report, Publications Office, 2018, https://data.europa.eu/doi/10.2759/781960

European Union Agency for Cybersecurity (ENISA) (2019): Pseudonymisation techniques and best practices according to the General Data Protection Regulation (GDPR). Online verfügbar unter https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices [Letzter Abruf 01.10.2024].

Farrall, F. et al. (2022): Data-sharing made easy, Deloitte, Online verfügbar unter https://www2.deloitte.com/content/dam/insights/articles/US165038_TT22-data-sharing/DI_TT22-Data-sharing.pdf. [Letzter Abruf 12.12.2024].

Fung, B.C M. et al. (2010): Privacy-Preserving Data Publishing. Online verfügbar unter https://charuaggarwal.net/generalsurvey.pdf [Letzter Abruf 07.11.2024].

Gonzales, A. et al. (2023): Synthetic data in health care: A narrative review. PLOS Digit Health 2(1), Online verfügbar unter https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082 [Letzter Abruf 16.12.2024].

Graef, I. et al. (2019): Limits and Enablers of Data Sharing An Analytical Framework for EU Competition – Data Protection and Consumer Law, Online Verfügbar https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3494212 [Letzter Abruf 12.12.2024].

Grunes, A. P./ Stucke, M. E. (2015): No Mistake About It: The Important Role of Antitrust in the Era of Big Data, Antitrust Source, University of Tennessee Legal Studies Research Paper No. 269 Online verfügbar unter https://ssrn.com/abstract=2600051 [Letzter Abruf 12.12.2024].

Hjørland, B. (2018): Data (with Big Data and Database Semantics). Knowledge Organization 45(8): 685-708. DOI:10.5771/0943-7444-2018-8-685

https://cloud.google.com/bigquery/docs/differential-privacy [Letzter Abruf 16.12.2024].

https://rechneronline.de/hash/sha256.php [Letzter Abruf 19.11.2024].

https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf [Letzter Abruf 16.12.2024].

https://www.ssldragon.com/de/blog/sha-256-algorithmus/#hashing-definition [Letzter Abruf 19.11.2024].

Independent European advisory body on data protection and privacy (2014): Opinion 05/2014 on Anonymisation Techniques. Online verfügbar unter https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [Letzter Abruf 28.10.2024].

Iyengar, J. (2002): Transforming Data to Satisfy Privacy Constraints, online verfügbar unter https://ci-teseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2676d77b4e4cc58250ed20b4f85576a9fb33ae5a [Letzter Abruf 26.11.2024].

Kellermann, J. & Gritzalis, D. (2020): Pseudonymization techniques and their use in distributed ledger technology environments to enhance privacy. Journal of Information Security and Applications, 53, 102518. https://doi.org/10.1016/j.jisa.2020.102518

Kohlmayer, F.; Lautenschläger, R. & Prasser, F. (2019): Pseudonymization for research data collection: is the juice worth the squeeze?, BMC Medical Informatics and Decision Making. Online verfügbar unter https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0905-x#:~:text=We%20conclude%20that%20(1)%20more%20research%20is%20needed

Li, N. et al. (2007): T-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Engineering. Online verfügbar unter https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf [Letzter Abruf 06.11.2024].

Machanavajjhala, A et al. (2007): L-Diversity: Privacy beyond k-anonymity, Online verfügbar https://d1wqtxts1xzle7.cloudfront.net/61486977/6l_Diversity_Privacy20191211-5159-1lpm4x-libre.pdf?1576087434=&response-content-disposition=inline%3B+file-name%3DDiversity_Privacy_Beyond_k_Anonymity.pdf&Expires=1732621064&Signature=WBxABC0S6Y9oSzV-TeqeRiB540m3T9087Z9W0HAA175y0dTajja2Y8TLd~DgC4b3f4xSldqLRZK0PpvT78CStf6AtPmmSNakpdh-iYf82TN7afsr-q~zUevI40FnPcG7awGZpXkaGL2yY8~n6AKxQH0uXZLtiXlYDlYi3aoX52ey5xSDk9Vfxjd-Nah9spZYWGnNnmZrHVXbBm6JnsrSV9FGw3OJCWzSyu7t~zWyKf8UocA1o2MmGTiu-KNdETukpMBTUYiOsxEUTZ3418ZTR0D08m2T50eYBRueRmfDF4piLbQl4pLR4MJ4-1CzHlyHsO1tK1~56Pfn1Dg9JbOM4UjVw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA [Letzter Abruf 26.11.2024].

Majeed, A. (2023): Attribute-Centric and Synthetic Data Based Privacy Preserving Methods: A Systematic Review. Online verfügbar unter https://www.mdpi.com/2624-800X/3/3/30 [Letzter Abruf 28.10.2024].

Majeed, A./ Lee, S. (2020): Anonymization Techniques for Privacy Preserving and Data Publishing: A Comprehensive Survey, IEEE Access. Online verfügbar unter https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9298747 [Letzter Abruf 07.11.2024].

Mivule, K. (2013): Utilizing Noise Addition for Data Privacy, an Overview. Online verfügbar unter https://arxiv.org/pdf/1309.3958 [Letzter Abruf 28.10.2024].

Narayanan, A./ Shmatikov, V. (2008: Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (S. 111–125). online verfügbar https://systems.cs.columbia.edu/private-systems-class/papers/Narayanan2008Robust.pdf [Letzter Abruf 12.12.2024].

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review, 57(6), 1701–1777. Online verfügbar: https://www.uclalawreview.org/broken-promises-of-privacy-responding-to-the-surprising-failure-of-anonymization [Letzter Abruf 12.12.2024].

Pfitzmann, A./ Hansen, M. (2010): A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. Online verfügbar unter https://dud.inf.tu-dresden.de/Anon_Terminology.shtml [Letzter Abruf 28.10.2024].

Rocher L. et al. (2019): Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 10: 3069

Rocher, L. et al. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, 10, 3069, online verfügbar https://www.nature.com/articles/s41467-019-10933-3 [Letzter Abruf 12.12.2024].

Salinas, S. O./ Lemus, A. C. (2017): Data warehouse and big data integration. Int. Journal of Comp. Sci. and Inf. Tech 9: 1-17, Online verfügbar unter https://d1wqtxts1xzle7.cloudfront.net/65085088/9217ijcsit01-libre.pdf?1606905610=&response-content-disposition=inline%3B+filename%3DData_Warehouse_and_Big_Data_Integration.pdf&Expires=1734044952&Signature=OKBz3wrRtPozQ8zvCOTqY99xwrGc6KWrT41OVxasIkT0Klk1Kj7QaV2Ubfqvrhb0lNNgoplqzlef-cfFgPRMEf7UTHNEdsqZHJyVfT-J04YmZDBR6f0Kf7CkISlE04rcyvyWeqI-lXqtUo~8XF3vEKqhlT0xvVSE3WxL09QZ7zi2YPLroqPn-8-XiM8Xwser92ckfRAlnhOuXyQ1jbcdN-s3oTCt6E2eC34HwPU-CtzKPg2SiFn-9PM1ANQgo51tiP73kzinmkpfCSDvM0qM14GzuXtdA4L~tkpKUWSTs1OR6HzxN~vstzb67H8XWEDmcMur4heh1xeqBK3adxg9o4Q__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA [Letzter Abruf 12.12.2024].

Samarati, P./ Sweeney, L. (1998): Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. Online verfügbar unter https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf [Letzter Abruf 28.10.2024].

Shapiro, C./ Varian, H.R. (2013): Information Rules: A Strategic Guide to the Network Economy: Harvard Business Press.

Slijepcevic, D. et al (2021): K-Anonymity in practice: How generalisation and suppression affect machine learning classifiers, Online verfügbar unter https://www.sciencedirect.com/science/article/pii/S0167404821003126 [Letzter Abruf 07.11.2024].

Stalla-Bourdillon, S. & Knight, A. (2017): Anonymous Data v. Personal Data – A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data (March 6, 2017). Wisconsin International Law Journal, 2017. Online verfügbar unter SSRN: https://ssrn.com/abstract=2927945

Stalla-Bourdillon, S./ da Rosa Lazarotto, B. (2024): Search queries and anonymisation: How to read Article 6(11) of the DMA and the GDPR together?, online verfügbar https://www.europeanlawblog.eu/pub/2uxr4anu/release/1 [Letzter Abruf 28.11.2024); Stalla-Bourdillon, S./Knight, A. (2017): Anonymous Data v. Personal Data — A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data (March 6, 2017). Wisconsin International Law Journal, online verfügbar https://ssrn.com/abstract=2927945

Sweeney, L. (2000): Simple Demographics Often Identify People Uniquely, online verfügbar unter https://dataprivacylab.org/projects/identifiability/paper1.pdf [Letzter Abruf 25.11.2024].

Sweeney, L. (2002a): Achieving k-Anonymity Privacy Protection using Generalization and Suppression, in International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 10, Issue 5, pp. 571-588. Online verfügbar unter https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.pdf#:~:text=anonymity%20provides%20privacy%20protection%20by%20guaranteeing%20that%20each [Letzter Abruf 02.10.2024].

Sweeney, L. (2002b): K-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.

Vardalachakis, M. et al. (2023): Anonymization, Hashing and Data Encryption Techniques: A Comparative Case Study, in: International Conference on Applied Mathematics and Computer Science (ICAMCS). IEEE, 2023. S. 129-135. Online verfügbar unter https://www.researchgate.net/profile/Marios-Vardalachakis/publication/378351505_Anonymization_Hashing_and_Data_Encryption_Techniques_A_Comparative_Case_Study/links/66b641c951aa0775f277a142/Anonymization-Hashing-and-Data-Encryption-Techniques-A-Comparative-Case-Study.pdf [Letzter Abruf 07.11.2024].

Voigt, P. & von dem Bussche, A. (2017): The EU General Data Protection Regulation (GDPR): A Practical Guide

Wang, K. & Li, J. (2005). Data Swapping: A Privacy Preserving Technique for Data Publishing. Proceedings of the 4th International Conference on Data Mining, ICDM 2004.

Wang, R. Y./ Strong, D. M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems, Vol. 12, No 4, pp.5-33. Online verfügbar unter http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf#:~:text=Richard%20Y.%20Wang%20is%20Associate%20Professor%20of%20Information [Letzter Abruf 30.09.2024].

ISSN 1865-8997