abida
ASSESSING BIG DATA

FÜR IMMER ANONYM: WIE KANN DE-ANONYMISIERUNG VERHINDERT WERDEN?

01IS15016A-F

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Datatree AG
Prof. Dr. Thomas Jäschke
Sina Rochow, Hanjo Tewes, Alexander Vogel, Henning Mertes

Baum, Reiter & Collegen
Prof. Dr. Julius Reiter, Dr. Olaf Methner

ABIDA - Assessing Big Data

PROJEKTLAUFZEIT 01.03.2015-28.02.2019

Westfälische Wilhelms-Universität Münster,
Institut für Informations-, Telekommunikations- und
Medienrecht (ITM), Zivilrechtliche Abteilung

Karlsruher Institut für Technologie,
Institut für Technikfolgenabschätzung
und Systemanalyse (ITAS)

Leibniz Universität Hannover
Institut für Rechtsinformatik
(IRI)

Technische Universität Dortmund,
Wirtschafts- und Sozialwissenschaftliche
Fakultät (WiSo) Techniksoziologie

Ludwig-Maximilians-Universität München,
Forschungsstelle für Information, Organisation
und Management (IOM)

Wissenschaftszentrum
Berlin für Sozialforschung

ABIDA - Assessing Big Data

Über das Gutachten

Das Gutachten wurde im Rahmen des ABIDA-Projekts mit Mitteln des Bundesministeriums für Bildung und Forschung erstellt. Der Inhalt des Gutachtens gibt ausschließlich die Auffassungen der Autoren wieder. Diese decken sich nicht automatisch mit denen des Ministeriums und/oder der einzelnen Projektpartner.

ABIDA lotet gesellschaftliche Chancen und Risiken der Erzeugung, Verknüpfung und Auswertung großer Datenmengen aus und entwirft Handlungsoptionen für Politik, Forschung und Entwicklung.

www.abida.de

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

BDSG	-	Bundesdatenschutzgesetz
BStatG	-	Bundesstatistikgesetz
BVerfG	-	Bundesverfassungsgericht
EU DS-GVO	-	Europäische Datenschutz-Grundverordnung
EuGH	-	Europäischer Gerichtshof
GG	-	Grundgesetz
IoT	-	Internet of Things

1 EINLEITUNG

Die Speicherung sehr großer Datenmengen ist heutzutage alltäglich. Daten werden zunehmend auch durch technische Geräte oder digitale Prozesse generiert und gespeichert. Schon früh wurde erkannt, dass mit den immer größer werdenden Kapazitäten der Massenspeicher und der Vernetzung von Computersystemen auch das Volumen der Informationen über einzelne Personen rapide ansteigt.[1]

Als Beispiele können Daten, die durch vernetzte Fahrzeuge generiert werden, ebenso dienen, wie Daten, die durch Smartphones erzeugt werden (Kommunikationsdaten, Bewegungsdaten, Daten in Bezug auf Internetnutzung, etc.). Auch im Bereich von Kunden-, Bank- und Kreditkarten werden viele Daten generiert. Weiterhin fallen überaus große Datenmengen im Bereich des Internet (z.B. Suchmaschinen, Social Media, etc.), in Industrie 4.0- und in IoT-Lösungen an.

Die Vorteile für Handel, Industrie und Forschung liegen auf der Hand: Sehr große Datenmengen führen zu wesentlich genaueren Ergebnissen bei Marktforschungen, es können beispielsweise energiesparende Verbrauchssteuerungen entwickelt werden, der Verlauf von Epidemien oder Unwettern kann besser prognostiziert werden u.v.m.

Big Data ist aber Fluch und Segen zugleich. Gerade im Bereich der Forschung ist es oft erforderlich, dass Daten veröffentlicht oder weitergegeben werden. Neben den o.g. Vorteilen besteht somit die Gefahr, dass anonyme sensible Informationen, die lediglich zu statistischen Zwecken erhoben wurden, später einzelnen Personen zugeordnet werden können und diese dadurch möglicherweise Nachteile erleiden. So könnten Informationen über tatsächlich vorhandene oder zukünftig potentiell auftretende Erkrankungen zu erhöhten Versicherungsbeiträgen führen.

Die einzige Möglichkeit, die Zuordnung von sensiblen Informationen zu einer Person zu verhindern, stellt die Anonymisierung der Daten dar. Allerdings ist ein anonymisierter Datenbestand kein Garant dafür, dass eine De-Anonymisierung nicht stattfinden kann. Werden die Daten nicht in ausreichendem Maße anonymisiert, besteht die Gefahr, dass - beispielsweise durch Hintergrundwissen oder im Zusammenhang mit weiteren Informationen - sensible Daten einer einzelnen Person zugeordnet werden können.

Das Ziel des Gutachtens ist es, Möglichkeiten zu eruieren, die eine dauerhafte Anonymisierung von Daten im Bereich Big Data sicherstellen und somit eine Zuordnung von sensiblen Informationen zu einer konkreten Person verhindern. Als Stakeholder sind in diesem Zusammenhang die betroffenen Personen, deren Daten verarbeitet werden, die Dateninhaber, die diese Daten sammeln und weitergeben, sowie die Datennutzer, die ein Interesse an der Auswertung solcher Daten haben, zu nennen.

Die Forschungsfrage des Gutachtens lautet daher:

Durch welche technischen und organisatorischen Maßnahmen kann De-Anonymisierung unter Berücksichtigung rechtlicher Aspekte und Anforderungen der einzelnen Stakeholder verhindert werden?

Nach den erläuternden Kapiteln werden im weiteren Verlauf des Gutachtens Anforderungen an technische und organisatorische Maßnahmen zur Anonymisierung eruiert und mit Hilfe einer Nutzwertanalyse auf Anwendbarkeit geprüft.

---

[1] (Sweeney 2001, S. 1)

Die Ausführungen in den Kapiteln 4.3.2, 5.6 und 5.7 sowie einzelne rechtliche Erläuterungen in den Kapiteln 2.2, 4.1, 4.3.1, Einleitung zu 5 und Teile des Fazits wurden verantwortlich von der Kanzlei Baum Reiter & Collegen geleistet.

Abschließend werden Handlungsempfehlungen zur Verhinderung von De-Anonymisierung auf Grundlage der Nutzwertanalyse gegeben.

2 GRUNDLAGEN DES DATENSCHUTZES

2.1 GRUNDSÄTZE DES DATENSCHUTZRECHTS

Das Datenschutzrecht ist ein Recht, das in der Vergangenheit insbesondere durch die Geltung von allgemeinen Prinzipen gekennzeichnet war, die bei der Auslegung von datenschutzrechtlichen Rechten und Pflichten eine Rolle spielten.[2] Diese Rechtssätze sind nun in Art. 5 EU DS-GVO kodifiziert und somit unmittelbar geltendes Recht.[3]

Rechtmäßigkeit der Datenerhebung

Der Art. 5 Abs. 1 lit. a Var. 1 EU DS-GVO beinhaltet den Grundsatz der Rechtmäßigkeit der Verarbeitung. Dieser Grundsatz besagt, dass eine Verarbeitung personenbezogener Daten grundsätzlich nur erfolgen darf, sofern diese rechtmäßig ist. Ob die Verarbeitung rechtmäßig ist, bestimmt sich danach, ob ein Rechtsgrund im Sinne des Art. 6 Abs. 1 EU DS-GVO vorliegt.

Treu und Glauben

Der Grundsatz von Treu und Glauben ist in Art. 5 Abs. 1 lit. a Var. 2 EU DS-GVO manifestiert. Eine Verarbeitung entspricht Treu und Glauben, wenn sie innerhalb dessen liegt, womit der Betroffene bei Zugrundelegung der rechtlichen Regeln redlicherweise rechnen muss.[4] Dadurch schützt der Grundsatz den Betroffenen davor, dass einmal erhobene Daten in einem Kontext verarbeitet werden, der für die betroffene Person nicht ersichtlich ist.

Transparenz

Der Grundsatz der Transparenz (Art. 5 Abs. 1 lit. a EU DS-GVO) ist an den Verantwortlichen der Datenverarbeitung dahingehend adressiert, dass er die personenbezogenen Daten nur in einer Weise verarbeitet, welche für die betroffene Person nachvollziehbar ist.[5] Dies bedeutet, dass die betroffene Person die Datenverarbeitung für die Vergangenheit nachvollziehen und für die Zukunft vorhersehen können muss.[6]

Zweckbindung

Damit die Person/Stelle, die die personenbezogenen Daten verarbeitet, nicht nach seinem/ihrem Belieben mit den Daten verfährt, besteht der Grundsatz der Zweckbindung. Er stellt sicher, dass die Rechtfertigung, die durch die Erhebung bzw. die Speicherung bestand, auch in einem späteren Kontext aufrecht erhalten bleibt.[7] Dieser Grundsatz ist in Art. 5 Abs. 1 lit. b EU DS-GVO niedergeschrieben.

Datenminimierung

---

2 (Vgl. Schantz in Schantz/Wolff 2017, Kap. D Rn. 380)
3 (Vgl. ebenda, Kap. D Rn. 381)
4 (Vgl. ebenda, Kap. D Rn. 393)
5 (Vgl. Thode in Schläger/Thode 2018, S. 35, Rn. 123)
6 (Vgl. ebenda)
7 (Vgl. Schantz in Schantz/Wolff 2017, Kap. D Rn. 408)

Der in Art. 5 Abs. 1 lit. c EU DS-GVO festgehaltenen Grundsatz der Datenminimierung bedeutet, dass die von dem Verantwortlichen erhobenen personenbezogenen Daten für den Zweck angemessen und erheblich sein müssen.[8] Darüber hinaus soll eine Beschränkung der Daten auf das für die Verarbeitungszwecke notwendige Maß beschränkt werden.[9]

Richtigkeit

Der Grundsatz der Richtigkeit aus Art. 5 Abs. 1 lit. d EU DS-GVO besagt, dass die personenbezogenen Daten, welche der Verantwortliche von der betroffenen Person verarbeitet, sachlich richtig sein müssen.[10] Diese Pflicht besteht aus dem Gedanken, dass in Zeiten der modernen Datenverarbeitung Daten nur schwer kontrolliert und noch schwerer korrigiert werden können.[11]

Speicherbegrenzung

Das Gebot der Speicherbegrenzung (Art. 5 Abs. 1 lit. e EU DS-GVO) enthält zwei Regelungsabsichten. Zum einen soll eine zeitliche Begrenzung der Speicherungsmöglichkeit vorgenommen werden, d.h. die Daten dürfen nur solange gespeichert werden, wie es zum Zweck der Erhebung erforderlich ist.[12] Darüber hinaus enthält es die Regelungsabsicht, die personenbezogenen Daten frühestmöglich zu anonymisieren.[13]

Integrität und Vertraulichkeit

Die Begriffe der Integrität und Vertraulichkeit sind in Art. 5 Abs. 1 lit. f EU DS-GVO geregelt. Der Begriff der Integrität beschreibt die Sicherstellung der Korrektheit (Unversehrtheit) von Daten und der korrekten Funktionsweise von Systemen.[14]

Vertraulichkeit ist der Schutz vor unbefugter Preisgabe von Informationen. Vertrauliche Daten und Informationen dürfen ausschließlich Befugten in einer zulässigen Weise zugänglich sein.[15]

Rechenschaftspflicht

Die Rechenschaftspflicht aus Art. 5 Abs. 2 EU DS-GVO verpflichtet den Verantwortlichen bei der Datenverarbeitung auf die Datenschutzgrundsätze, die in Art. 5 Abs. 1 EU DS-GVO aufgeführt sind. Durch diese Norm hat der für die Verarbeitung Verantwortliche eine Darlegungslast, dass die zuvor genannten Grundsätze durch ihn eingehalten worden sind.[16]

2.2 RECHTE DER BETROFFENEN

Die EU DS-GVO räumt den Betroffenen Rechte ein, die sie bezüglich ihrer personenbezogenen Daten geltend machen können. Diese sind primär in den Vorschriften des Art. 12-23 EU DS-GVO geregelt.

---

8 (Vgl. Thode in Schläger/Thode 2018, S. 37, Rn. 130)
9 (Vgl. ebenda)
10 (Vgl. Schantz in Schantz/Wolff 2017, Kap. D Rn. 441)
11 (Vgl. ebenda, Kap. D Rn. 444)
12 (Vgl. ebenda)
13 (Vgl. ebenda, Kap. D Rn. 380)
14 (Vgl. Online-Glossar des Bundesamtes für Sicherheit in der Informationstechnik)
15 (Vgl. Online-Glossar des Bundesamtes für Sicherheit in der Informationstechnik)
16 (Vgl. Schantz in Schantz/Wolff 2017, Kap. D Rn. 449)

Auskunftsrecht

Der Art. 15 Abs. 1 EU DS-GVO räumt Personen das Recht ein, Auskunft zu verlangen. Aufgrund dieser Norm sind Betroffene in der Lage sich an die Stellen zu wenden, von der sie wissen oder vermuten, dass diese personenbezogenen Daten von dem Betroffenen verarbeiten. Das Auskunftsrecht des Art. 15 Abs. 1 EU DS-GVO ist weit gefasst. Die verarbeitende Stelle muss, sofern sie personenbezogene Daten von dem Betroffenen verarbeitet oder verarbeitet hat, über die in Art. 15 Abs. 1 a-h EU DS-GVO aufgeführten Punkte Auskunft erteilen. Darüber hinaus ist sie verpflichtet, den Betroffenen eine kostenlose erste Kopie der personenbezogenen Daten, welche die Stelle von dem Betroffenen verarbeitet/verarbeitet hat, zur Verfügung zu stellen.[17]

Berichtigungsrecht

In Art. 16 EU DS-GVO ist geregelt, dass eine betroffene Person unverzüglich von dem Verantwortlichen eine Berichtigung von unrichtigen personenbezogenen Daten verlangen kann. Ferner kann die betroffene Person die Vervollständigung von unvollständigen Daten verlangen. Dieses Recht der betroffenen Person ist Ausdruck des Grundsatzes der Richtigkeit und gibt den betroffenen Personen die Möglichkeit, diese im Zweifelsfall wiederherzustellen.

Löschung

Der Betroffene hat auch das Recht, seine personenbezogenen Daten löschen zu lassen (Art. 17 EU DS-GVO). Dieses Recht kann er geltend machen, sofern die in Art. 17 Abs. 1 EU DS-GVO aufgeführten Gründe vorliegen. Ferner dürfen keine Gründe im Sinne des Art. 17 Abs. 3 EU DS-GVO vorliegen, aus denen der Verantwortliche nicht zu einer Löschung verpflichtet ist.

Einschränkung der Verarbeitung

Bei der Einschränkung der Verarbeitung nach Art. 18 EU DS-GVO besteht für die betroffene Person die Möglichkeit, eine Verarbeitung der personenbezogenen Daten zu beschränken. Hierdurch soll sichergestellt werden, dass die Daten nur noch für bestimmte Zwecke verarbeitet werden.[18]

Datenübertragbarkeit

Der Art. 20 Abs. 1 EU DS-GVO räumt der betroffenen Person die Möglichkeit ein, ihre personenbezogenen Daten von der erhebenden Stelle in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten.

Der Art. 20 Abs. 2 EU DS-GVO regelt sogar, dass erwirkt werden kann, dass die Daten von einem Verantwortlichen zu einem neuen Verantwortlichen übermittelt werden.

Mit dieser Vorschrift möchte man erreichen, dass in einem Zeitalter in dem Daten immer wichtiger werden, eine Einspeisung dieser personenbezogenen Daten in die Systeme andere Anbieter sichergestellt wird. Mit dieser Portabilität ist eine freie Anbieterwahl auch dann gewährleistet, wenn das Produkt bzw. seine Qualität maßgeblich von den bisher erhobenen und ausgewerteten Daten abhängen. Als Beispiel kann man hier z.B. Trainingscomputer anführen, die sich der Leistungsentwicklung der Betroffenen anpassen und das Trainingspensum an dessen Zielen orientieren.

---

17 (Vgl. Art. 15 Abs. 3 DSGVO)

18 (Vgl. Herbst in Kühling/Büchner 2018, Art. 18 Rn. 29)

Widerspruch

Der Art 21 Abs. 1 EU DS-GVO räumt der betroffenen Person das Recht auf Widerspruch ein, darunter wird verstanden, dass die betroffene Person bei der Verarbeitung personenbezogener Daten aufgrund eines gesetzlichen Erlaubnistatbestandes widersprechen kann.[19] Dieser Erlaubnistatbestand ist in der Regel der des überwiegenden berechtigten Interesses und führt meist zu einem Verbot der Verarbeitung.[20] Dies gilt jedoch ausnahmsweise nicht, sofern der Verantwortliche schutzwürdige Gründe für die Verarbeitung nachweisen kann.[21]

Anhand der Definition von personenbezogenen Daten und der identifizierbaren natürlichen Person in Art. 4 Abs. 1 EU DS-GVO sowie des Begriffs des Verarbeitens ist ersichtlich, dass es sich bei fast allen Daten um personenbezogene Daten handelt. Für diese gelten grundsätzlich die Regeln des Datenschutzes. Dies bedeutet aber auch, dass die Kombination von Daten fast immer dazu führt, dass eine Person identifizierbar ist. Dies ist insbesondere unter den neuen Methoden der Datenerhebung und Verarbeitung problematisch, da die Auswertung von Daten Rückschlüsse auf die Person zulässt und zu einer Kollision mit dem Datenschutz führen kann. Im Oktober 2016 entschied der EuGH, dass Informationen ohne direkten Personenbezug im Besitz der Stelle, die rechtmäßig Zugriff auf ausreichende Zusatzdaten hat, um die Informationen mit einer Person zu verknüpfen und diese damit zu identifizieren (allerdings nur diese Stelle), als personenbezogene Daten angesehen werden können. Damit Daten als personenbezogene Daten behandelt werden, reiche es aus, dass die verantwortliche Stelle die ihr vernünftigerweise zur Verfügung stehenden rechtlichen Mittel einsetzen könne, um von einem Dritten entsprechendes Zusatzwissen zu erhalten, durch das sie die jeweilige Person identifizieren kann.

---

19 (Vgl. Thode in Schläger/Thode 2018, S. 47, Rn. 179)
20 (Vgl. ebenda, S. 47, Rn. 179 ff.)
21 (Vgl. ebenda, S. 48, Rn. 181)

3 BIG DATA

Daten werden häufig als Öl oder Gold des 21. Jahrhunderts tituliert.[22] Man kann diese These bestätigt sehen, wenn man den Börsengang des Unternehmens Facebook[23] im Jahr 2012 genauer betrachtet. Ein Unternehmen, welches zum damaligen Zeitpunkt 1 Milliarden Dollar pro Jahr verdiente, wurde mit einem Firmenwert von 104 Milliarden Dollar bewertet.[24] Viele sehen den Wert des Unternehmens in seinen inzwischen 2,2 Milliarden[25] monatlich aktiven Nutzern. Es stellt sich hier die Frage, weshalb die Daten einen solch hohen Wert haben sollen.

Durch Big Data ergeben sich inzwischen andere Möglichkeiten, vorhandene Daten auszuwerten und zu nutzen. Auch wenn die Bewertung hierdurch nur teilweise erklärt wird, liefert es dennoch Erklärungsansätze, warum viele (teils auch zusammenhangslose) Daten dennoch ein enormes wirtschaftliches Potenzial beinhalten.

3.1 BEGRIFFBESTIMMUNG BIG DATA

Es gibt viele Versuche in der Literatur, den Begriff Big Data zu definieren.[26] Trotz vielfältiger Ansätze hat sich keine Definition durchsetzen können. Sofern man jedoch nach der Definition sucht, die bis heute am gängigsten ist[27], gelangt man zu den Definitionen der „V´s" die den Grundcharakter von Big Data ausmachen.[28]

3.1.1 DIE DEFINITION DER DREI V'S

Eine der gängigsten Definitionen verortet in Big Data die drei Eigenschaften „Volume", „Velocity" und „Variety".[29]

Volume

Der Begriff „Volume" bezeichnet die Datenmenge, die aufgrund von Big Data Methoden verarbeitet werden können.[30] Diese Datenmenge kann im Terabyte- bis Zetabytebereich liegen.[31]

In diesem Kontext spielt das Moore´sche Gesetz eine besondere Rolle.[32] Das Moore´sche Gesetz besagt vereinfacht, dass sich die Rechenleistung innerhalb eines bestimmten Zeitraums (je nach

---

22 (Vgl. Märmecke et al. in Märmecke/Passoth/Wehner 2018, S. 1) und auch (Vgl. Schwanebeck 2017, S. 14)
23 Facebook sollte exemplarisch für Unternehmen betrachtet werden, welche viele Daten ihrer Nutzer gespeichert haben.
24 (Vgl. https://www.zeit.de/wirtschaft/geldanlage/2012-05/facebook-nashdaq-handelsstart)
25 (Vgl. https://www.zeit.de/news/2018-04/25/facebook-meldet-weiter-starke-zahlen-nach-datenskandal-180425-99-54728)
26 (Vgl. King 2014, S. 34 ff.)
27 (Vgl. Dorschel 2015, S. 6)
28 (Vgl. King 2014, S. 35)
29 (Vgl. Fasel und Meier 2016, S. 5 f.)
30 (Vgl. Desoi 2018, S. 14)
31 (Vgl. Fasel und Meier 2016, S. 6)
32 (Vgl. Dorschel 2015, S. 7)

Interpretation alle 12 bis 24 Monate) verdoppelt.[33] Durch diese Konstellation nimmt der Grad der Vernetzung zu, indem Geräte und Alltagsgegenstände bei menschlichen Tätigkeiten unterstützen und Daten erheben.[34] Das exponentielle Wachstum an Rechenleistung steht in unmittelbarer Relation dazu, dass sich etwa alle zwei Jahre das weltweite Datenvolumen verdoppelt.[35]

Diese Masse an neuen Daten gilt als Grundlage von Big Data, da immer mehr Daten aus allen Lebensbereichen zur Verfügung stehen, die analysiert werden können.[36]

Velocity

Für den Begriff der „Velocity" gibt es unterschiedliche Deutungsarten. Teilweise wird hierunter die Geschwindigkeit verstanden, mit der Daten produziert und verändert werden, da dies eine rasche Analyse und Entscheidungsfindung verlangt.[37] Eine andere Ansicht definiert den Begriff dadurch, dass der Zeitraum in den Daten anfallen immer schneller, die Zeit, in der die Daten jedoch wertschöpfend sind, immer kürzer wird.[38]

Unabhängig davon, welcher dieser Definitionen man folgt, besteht zwischen „Velocity" und „Volume" eine unmittelbare Wechselbeziehung, denn je schneller die Daten errechnet werden, desto mehr Daten werden in einer immer kürzeren Zeit hergestellt.[39]

Variety

Unter der „Variety" versteht man die Vielzahl von Datenquellen, welche nicht nur auf eine immer größere Datenanzahl zurückzuführen ist, sondern auch die Möglichkeit auf eine Menge verschiedenartiger Datenquellen zurückgreifen zu können.[40] Diese verschiedenartigen Datenquellen können zum Beispiel strukturierte, teilstrukturierte oder unstrukturierte Datenbanken sein.[41] Aber nicht nur Datenbanken kommen zur Auswertung in Betracht, eine Auswertung kann beispielsweise auch von Social Media-Inhalten,[42] Videodateien, Audiodateien oder Messdaten erfolgen.[43]

3.1.2 DIE DEFINITION DER VIER V'S

Häufig wird zusätzlich zu den vorherigen 3 V´s „Veracity" als viertes Kriterium zur Definition von Big Data herangezogen.[44]

In der deutschen Sprache wird dies häufig mit „Wahrhaftigkeit" übersetzt. Dieses Kriterium zielt darauf ab, dass die zur Analyse vorgesehenen Daten häufig nicht solchen Ursprungs sind, dass man

---

33 (Vgl. ebenda)
34 (Vgl. Dorschel 2015, S. 7)
35 (Vgl. ebenda m.w.N.)
36 (Vgl. ebenda, S. 7)
37 (Vgl. King 2014, S. 35)
38 (Vgl. Fasel 2014, S. 389)
39 (Vgl. Dorschel 2015, S. 7)
40 (Vgl. Desoi 2018, S. 14)
41 (Vgl. ebenda)
42 (Vgl. Dorschel 2015, S. 8) und (Vgl. Desoi 2018, S. 14)
43 (Vgl. Desoi 2018, S. 14)
44 (Vgl. IBM, http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data)

unbeschränkt von der Richtigkeit der Daten ausgehen kann.[45] Auch die Vollständigkeit ist nicht immer gegeben.[46]

Exemplarisch kann man dies an Social Media-Daten erläutern. Texte, die Mitglieder auf einer Social Media-Plattform schreiben, sind meist subjektiv und stammen je nach User und Eintrag aus einem zeitlich und inhaltlich unterschiedlichen Kontext.[47] Aus diesem Grund ist es bei der Nutzung von Big Data eine der zentralen Aufgaben, diese Faktoren bei der Planung, Durchführung und Bewertung von Analysen zu berücksichtigen.[48]

3.1.3 KRITIK AN DER DEFINITION DER V'S

An den Definitionen der V´s wird kritisiert, dass sie zu informatikorientiert seien,[49] ferner wird angeführt, dass nach anderthalb Jahrzehnten inflationärer Nutzung von V-Worten, immer noch nicht deutlich sei, was Big Data im Kern bedeutet.[50] So führt Rust z.B. aus „Definitionsversuche sind im Einzelnen unzureichend, insgesamt vage, weil sie eine Menge an oft nicht kompatiblen Möglichkeiten für Voraussetzungen und Anwendungen der Datengenese, Datensammlung und Datenanalyse und schließlich Dateninterpretation umfassen.“[51]

Eine einheitlich anerkannte Definition des Begriffes Big Data wird es wohl auch in Zukunft nicht geben. Teilweise ist auch ein gegenläufiger Trend zu erkennen. So führt Dorschel aus, dass die Frage, was Big Data sei, in der Regel unzureichend und damit unzutreffend beantwortet würde, da dieser Frage der Vielschichtigkeit des Themas entgegenstehe.[52] Vielmehr benennt er, dass unter dem Begriff Big Data eine Vielzahl von Technologien, analytischer Methoden, Modellierungs- und Designverfahren, kommerzieller Konzepte sowie rechtlichen Rahmenbedingungen subsumiert werden können.[53] Somit macht er klar, dass die Suche nach einer Definition nicht allumfassend zu erreichen ist, da der Begriff inzwischen zu vielschichtig sei.

Zu diesem gegenläufigen Trend gehört es auch, dass um den Begriff Big Data bzw. Teilaspekte zu spezifizieren, gelegentlich konkurrierende Begriffe verwendet werden, die einen spezifischen Fokus von Big Data betonen: “Smart Data“ betont beispielsweise die sinnvolle Datenverwendung oder „Fast Data“ die Geschwindigkeit der Datenverarbeitung.[54]

Eine allgemein anerkannte und allumfassende Definition des Begriffes Big Data gibt es zum momentanen Zeitpunkt nicht. Sofern es ihn irgendwann geben sollte, wird er sicherlich nicht in den allgemeinen Sprachgebrauch einziehen und von jeder Person problemlos verstanden und verwendet werden können.

---

45 (Vgl. Desoi 2018, S. 14)
46 (Vgl. Dorschel 2015, S. 8)
47 (Vgl. ebenda)
48 (Vgl. ebenda, S. 8)
49 (Vgl. ebenda, S. 6)
50 (Vgl. Rust 2017, S. 10)
51 (Vgl. ebenda)
52 (Vgl. Dorschel 2015, S. 1)
53 (Vgl. ebenda, S. 2 ff.)
54 (Vgl. Gadatsch und Landrock 2017, S. 2)

Wie kann man den Begriff Big Data also massenkompatibel (wenn auch nicht vollständig) definieren? Für das allgemeine Verständnis von Big Data, erscheint die folgende Aussage daher passend:

Eine Kernaufgabe von Big Data ist es, übergreifende Kenntnisse aus unterschiedlichen Quellen und Formaten zu gewinnen.[55] Das Ziel ist, dass Computer die Dateninhalte selber verstehen und in der Lage sind, diese zu interpretieren.[56]

Die Möglichkeiten von Big Data sind evident. Es werden inzwischen sehr große Mengen von neu entstandenen Daten nach Mustern durchsucht, deren Masse der Mensch allein niemals bewältigen könnte. Ferner sind inzwischen nur Rechensysteme bei diesen Datenmengen in der Lage, Muster bzw. Zusammenhänge zu erkennen, die bei einer menschlichen Betrachtung nicht offensichtlich sind.

3.2 RISIKEN VON BIG DATA

Durch die in Big Data Systemen immer vorliegenden sehr hohen Rechenleistungen und den Zugriff auf viele unterschiedliche Quellen durch solche Systeme ist die Gefahr hoch, dass eigentlich anonyme Daten durch De-Anonymisierung einer Person zugeordnet werden können. Durch Auswertungen in Big Data Systemen wird „das Konzept der Anonymisierung, wenn nicht generell, so doch in vielen bisher als anonym bewerteten Situationen in Frage“[57] gestellt.

Inzwischen sind viele Möglichkeiten der Sammlung und Analyse der Daten vorhanden. Inzwischen wird durch Internet Tracking unser Surf-, Nutzungs- und Kommunikationsverhalten analysiert wodurch wir im Hinblick auf unser Konsumverhalten berechenbar und vorhersehbar sind.[58] Anhand wiederkehrender Zahlungen mit Bank- und Kreditkarten können Personen mit einer hohen Wahrscheinlichkeit identifiziert werden.[59] Erste Krankenkassen bieten Rabattsystem an, falls man sein Verhalten per Fitnessarmband analysieren lässt.[60] Daten werden also bereits in sehr hohem Umfang erhoben. Die Zusammenführung und Auswertung dieser Daten kann potentiell zu Nachteilen bei den betroffenen Personen führen.

Das Risiko, das durch Big Data entstehen könnte, soll an einem rein fiktiven Beispiel verdeutlicht werden.

Daten werden mittlerweile in fast jedem Lebensbereich erhoben und gespeichert:

- Bei der Bezahlung per EC-Karte oder Kreditkarte werden die Informationen zu den entsprechenden Transaktionen beim entsprechenden Abrechnungsdienstleister gespeichert.
- Bei der Nutzung von Rabattsystem wie beispielsweise Payback werden ebenfalls Daten über die Transaktion beim entsprechenden Anbieter gespeichert.
- Daten über Bewegungen werden beim Mobilfunkdienstleister gespeichert.
- Suchverläufe werden von Suchmaschinenanbietern gespeichert und zur Generierung von Profilen genutzt.

---

55 (Vgl. Dorschel 2015, S. 8)
56 (ebenda)
57 (Vgl. Roßnagel et al. 2016, S. 27)
58 (Weichert 2013, S. 255)
59 (Vgl. Kühl 2015)
60 (Vgl. Saft 2017)

- Die von Fitnessarmbändern generierten Daten werden in der Cloud gespeichert und teilweise sogar freiwillig an Krankenkassen übermittelt.

Schon wenige Transaktionen reichen aus, um Personen mit einer sehr hohen Wahrscheinlichkeit identifizieren zu können. Forscher vom MIT und der Universität Aarhus haben in einer Studie belegt, dass die Informationen über Ort und Zeitpunkt von lediglich vier Transaktionen ausreichen, um 90% der Personen in einer anonymisierten Liste zu identifizieren.[61]

Die Kombination der oben aufgeführten Daten ermöglicht eine wesentlich bessere Zuordnung von Merkmalen zu einer bestimmten Person, vor allem in der Kombination mit weiteren Daten (z.B. erworbene Adressdaten). Problematisch kann hier u.U. eine Fehlinterpretation durch Algorithmen sein, die möglicherweise Personen aus statistischer Sicht einer bestimmten Gruppe zuweisen. Diese Zuweisung kann ein Faktor in entscheidungsunterstützenden oder Prognose-Systemen genutzt werden.

Je mehr Datenquellen zur Verfügung stehen, desto weniger Einfluss hat die einzelne Person, durch ihren persönlichen Lebenswandel auf diese Wahrscheinlichkeiten Einfluss zu nehmen. Dies könnte bedeuten, dass in Zukunft Menschen in allen Lebensbereichen Nachteile erleiden könnten, weil sie statistisch einer Gruppe zugehörig erscheinen, der Sie aber in Wirklichkeit nicht angehören. Hierdurch könnte eine „Machtlosigkeit" entstehen, da wir als Individuen nicht in der Lage sind, unser Leben zu gestalten und zu beeinflussen, weil dieser Gestaltung Daten entgegenstehen, die wir als einzelne Person kaum beeinflussen können.

Wie die bisherigen Ausführungen zeigen, sind Big Data-Anwendungen in der Lage, große Mengen an Daten aus unterschiedlichen Quellen zu analysieren und den Inhalt in einen Kontext zu bringen. Der Datenschutz wiederrum hat die Intention, personenbezogene Daten zu schützen. Genau in dieser Schnittstelle liegt die Problematik des Datenschutzes, welcher potentiell mit Big Data kollidieren kann.

Dies soll an einem weiteren Beispiel verdeutlicht werden.

Netflix war im Jahr 2007 noch ein Online-DVD-Verleih und wollte die Empfehlungen, die sie für Filme aufgrund des Leihverhaltens ihrer Kunden aussprach, verbessern.[62] Aus diesem Grund anonymisierte Netflix 100 Millionen Filmbewertungen von 500.000 Netflix-Kunden und stellte sie Programmierern zur Verfügung, damit sie die Empfehlung weiterer Filme, verbessern konnten.[63]

Durch die vorgenommene Anonymisierung hatten die Daten auf den ersten Blick kein personenbezogenes Datum mehr.

Diese von Netflix bereitgestellten Daten konnten zwei Forscher der University of Texas in Austin jedoch einigen bestimmten Personen zuordnen.[64] Hierzu glichen sie die Filmkritiken die Netflix zur Verfügung gestellt hatte mit einem öffentlich zugänglichen Filmbewertungsportal ab, bei welchem viele Rezensenten ihren Klarnamen verwendeten.[65] Anhand von weniger geläufigen Filmen und deren Bewertungen war es möglich die Netflix-Kunden mit einer Wahrscheinlichkeit von 84% zu identifizieren.[66] Sofern auf das Datum der Filmbewertung zurückgegriffen werden konnte, stieg die Identifizierungswahrscheinlichkeit auf 99%. Darüber hinaus war es teilweise möglich, aus den

---

61 (Vgl. de Montjoye 2015, S. 5)
62 (Vgl. Hornung und Herfurth 2018, S. 164 f. m.w.N.)
63 (Vgl. ebenda)
64 (Vgl. ebenda)
65 (Vgl. ebenda)
66 (Vgl. ebenda)

Filmbewertungen Informationen wie z.B. die Religionszugehörigkeit, die politische Überzeugung oder die Sexualität abzuleiten.[67]

Dieser Vorfall zeigt, dass sofern man zunächst unabhängige Datenquellen kombiniert, ein anonymisierter Datensatz durch das Hinzufügen neuer Daten aus anderen Quellen sehr schnell de-anonymisiert werden kann.

Das Beispiel zeigt eindrucksvoll, dass die Personen, sofern sie im Internet einen Klarnamen bei Bewertungen hinterlassen, damit rechnen mussten, identifiziert zu werden. Die Intention dieses Beispiels ist es aufzuzeigen, mit welchen profanen Mitteln die De-Anonymisierung von Personen möglich war.

Sofern man berücksichtigt, dass automatisiert inzwischen tausende von Merkmalen abgeglichen werden können und der potentiell anonymisierte Datensatz keine Filmbewertung ist, sondern z.B. ein Krankheitsverlauf, eine sexuelle Vorliebe oder aber andere Dinge, die man nicht unbedingt fremden Personen erzählen würde, sollte dies nachdenklich stimmen.

---

67 (Vgl. ebenda)

4 PSEUDONYMISIERUNG UND ANONYMISIERUNG

Zwischen der Pseudonymisierung und der Anonymisierung von Daten kann wie folgt unterschieden werden:

Bei der Pseudonymisierung von Daten besteht die Möglichkeit, dass eine konkrete Person unter Hinzuziehung von gesondert aufbewahrten Informationen oder durch Zuordnungstabellen wieder identifiziert werden kann. Die Zuordnung von Informationen zu einer Person ist gewollt.

Bei der Anonymisierung von Daten können die betroffenen Personen nicht oder nur mit verhältnismäßig großem Aufwand wieder identifiziert werden.[68] Die Zuordnung von Informationen zu einer Person ist nicht gewollt.

4.1 PSEUDONYMISIERUNG

Im Allgemeinen kann Pseudonymisierung als eine Schutzmaßnahme für die Verarbeitung von Daten durch Dritte bezeichnet werden. Dabei soll gegenüber Dritten der Personenbezug der Daten ausgeschlossen werden. Laut § 46 Nr. 5 BDSG-neu bedeutet Pseudonymisierung „die Verarbeitung personenbezogener Daten in einer Weise, in der die Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die Daten keiner betroffenen Person zugewiesen werden können." Dies entspricht im Wesentlichen auch der Definition in Art. 4 Nr. 5 EU DS-GVO.

Es muss also bei der Pseudonymisierung zunächst gewährleistet sein, dass ohne Hinzuziehung zusätzlicher Informationen keine Zuordnung der Daten zu einer spezifischen Person möglich ist. Diese „zusätzliche Information" stellt eine Zuordnungsregel dar, mit der die pseudonymisierten Datensätze und die zugehörigen Personen wieder zusammengeführt werden könnten.

Weiterhin müssen die zusätzlichen personenbezogenen Informationen getrennt aufbewahrt werden, sodass nicht ohne weiteres eine Zusammenführung der Daten erfolgen kann. Eine logische Trennung mit unterschiedlichen Zugriffsberechtigungen kann dabei zum Einsatz kommen. Zudem muss im Vorhinein geklärt werden, welche Personen über bestimmte Zuordnungstabellen oder Verschlüsselungsverfahren verfügen und wer das Pseudonym generiert. Durch technische und organisatorische Maßnahmen muss also sichergestellt werden, dass eine unbefugte Re-Identifikation nicht erfolgen kann. Da der Verantwortliche immer Kenntnis des Zuordnungsschlüssels hat oder ihm diese Kenntnis (z.B. innerhalb eines Unternehmens) zugerechnet wird, müsste der Zuordnungsschlüssel bei einem externen Datentreuhänder aufbewahrt werden. Nach Erwägungsgrund 29 der EU DS-GVO ist es ausreichend, wenn die Beschränkung des Zugriffs auf den Zuordnungsschlüssel beim Verantwortlichen sichergestellt wird.

Der Unterschied zwischen der Anonymisierung und der Pseudonymisierung liegt darin, dass bei einer Pseudonymisierung der Personenbezug möglicherweise wieder hergestellt werden muss oder soll,

---

[68] (Vgl. Plath und Schreiber in Plath 2018, Rn. 58)

wohingegen bei der Anonymisierung der Personenbezug dauerhaft unterbleiben soll. Ein Beispiel für die Pseudonymisierung ist die Teilnahme an einer medizinischen Studie. Hier ist die Zuordnung der erhobenen Daten zu den individuellen Teilnehmern auszuschließen, kann aber später wieder erforderlich werden, um z.B. die Teilnehmer über unerwartete Nebenwirkungen eines getesteten Medikaments zu informieren.

Auch wenn der Personenbezug von pseudonymisierten Daten aufgrund der zwischengeschalteten Zuordnungsregel eingeschränkt ist, handelt es sich hier weiterhin im Grundsatz um personenbezogene Daten. Dies ist aus Erwägungsgrund 26 S. 2 EU DS-GVO ersichtlich. Die Daten unterfallen somit vollständig dem Anwendungsbereich der EU DS-GVO und werden dort in Art. 6 Abs. 4 lit. e), Art. 25 Abs. 1, Art. 32 Abs. 1 lit. a), Art. 40 Abs. 2 lit. d) und Art. 89 behandelt, wobei die Pseudonymisierung v.a. als Schutzmechanismus verstanden wird.

Im Allgemeinen kann Pseudonymisierung als eine Schutzmaßnahme für die Verarbeitung von Daten durch Dritte bezeichnet werden. Dabei soll gegenüber Dritten der Personenbezug der Daten ausgeschlossen werden.[69]

Dazu zählt ebenfalls, dass durch die technischen und organisatorischen Maßnahmen eine Zuordnung von personenbezogenen Daten durch weitere herangezogene Informationen auszuschließen ist.[70]

Abbildung 4-1 zeigt eine nicht ausreichende Pseudonymisierung. Im Beispiel ist zu erkennen, dass das Pseudonym nicht zufällig generiert wurde, sondern aus den Anfangsbuchstaben des Vor- und des Nachnamens besteht. Zudem sind die Straße und das Geburtsdatum in der Faktentabelle vorhanden, sodass mit ein wenig Hintergrundwissen eine eindeutige Zuordnung der Informationen zu einer Person möglich ist.

PersAuswNr	Vorname	Nachname	Straße	PLZ	Ort	Geburtsdatum	Beruf	Geschlecht	Kredit
918273645	Max	Muster	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	m	20000

Zuordnungstabelle

Pseudonym	PersAuswNr	Vorname	Nachname
MaMu	918273645	Max	Muster

Faktentabelle

Pseudonym	Straße	PLZ	Ort	Geburtsdatum	Beruf	Kredit	Geschlecht
MaMu	Teststraße 234	98765	Musterort	01.01.1960	Arzthelfer	20000	m

Abbildung 4-1: unzureichende Pseudonymisierung (Beispiel 1)

In Abbildung 4-2 wurde zwar das Pseudonym als zufälliger Wert generiert, dennoch ist weiterhin eine eindeutige Zuordnung der Informationen zu einer Person mit einer hohen Wahrscheinlichkeit möglich, wenn der Angreifer mit Hintergrundwissen ausgestattet ist.

PersAuswNr	Vorname	Nachname	Straße	PLZ	Ort	Geburtsdatum	Beruf	Geschlecht	Kredit
918273645	Max	Muster	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	m	20000

Zuordnungstabelle

Pseudonym	PersAuswNr	Vorname	Nachname
ID975312864	918273645	Max	Muster

Faktentabelle

Pseudonym	Straße	PLZ	Ort	Geburtsdatum	Beruf	Kredit	Geschlecht
ID975312864	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	20000	m

Abbildung 4-2: unzureichende Pseudonymisierung (Beispiel 2)

69 (Vgl. Knopp 2015, S. 527)
70 (Vgl. Schwartmann und Weiß 2017, S. 10)

In Abbildung 4-3 liegt eine korrekt pseudonymisierte Faktentabelle vor, da hier die Informationen zur Straße entfernt wurde und zudem die PLZ durch einen PLZ-Bereich und das Geburtsdatum durch das Alter ersetzt wurden. Eine Zuordnung der Informationen zu einer Person ist aber dennoch weiterhin möglich, da durch das Alter in Kombination mit dem PLZ-Bereich und dem Beruf und dem Geschlecht eine relativ kleine Menge an möglichen Individuen existiert.

PersAuswNr	Vorname	Nachname	Straße	PLZ	Ort	Geburtsdatum	Beruf	Geschlecht	Kredit
918273645	Max	Muster	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	m	20000

Zuordnungstabelle

Pseudonym	PersAuswNr	Vorname	Nachname
ID975312864	918273645	Max	Muster

Faktentabelle

Pseudonym	PLZ-Bereich	Alter	Beruf	Kredit	Geschlecht
ID975312864	98*	58	Arzthelfer	20000	m

Abbildung 4-3: ausreichende Pseudonymisierung

In Abbildung 4-4 wird durch die Bildung einer Altersgruppe von 50 bis 59 Jahren und eine Einordnung in eine Berufsgruppe eine effektive Pseudonymisierung erreicht. Die theoretische Menge an möglichen Individuen ist hier so groß, dass eine De-Pseudonymisierung nicht ohne weiteres möglich ist.

PersAuswNr	Vorname	Nachname	Straße	PLZ	Ort	Geburtsdatum	Beruf	Geschlecht	Kredit
918273645	Max	Muster	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	m	20000

Zuordnungstabelle

Pseudonym	PersAuswNr	Vorname	Nachname
ID975312864	918273645	Max	Muster

Faktentabelle

Pseudonym	PLZ-Bereich	Altersgruppe	Berufsfeld	Kredit	Geschlecht
ID975312864	98*	50 bis 59	Medizin	20000	m

Abbildung 4-4: effektive Pseudonymisierung

Wie oben beschrieben, ist ein Kennzeichen für Pseudonymisierung, dass der Personenbezug von Daten möglicherweise wiederhergestellt werden muss oder sogar soll. So ist zum Beispiel die Auflösung eines Pseudonyms im Falle einer Strafverfolgung auf Anforderung durch einen Staatsanwalt gefordert, um einen Täter identifizieren zu können.

4.2 ANFORDERUNGEN AN PSEUDONYME

Bei einer Verarbeitung von pseudonymisierten Daten, kann es vorkommen, dass bestimmte Inhalte/enthaltene Informationen des pseudonymisierten Datensatzes genutzt werden. Dazu müssen die vorher festgelegten Eigenschaften der ansonsten verdeckten Informationen verfügbar gemacht werden. Diese werden als Pseudonyme mit Verfügbarkeitsoptionen bezeichnet. Somit wird durch die Pseudonymisierung die Vertraulichkeit der zu schützenden Daten gewährleistet und zusätzlich kann durch die Verfügbarkeitsoption der Pseudonyme ein gewisses Maß an Nutzbarkeit bereitgestellt werden.

Eine Verfügbarkeitsoption ist unter anderem die Rückführung eines Pseudonyms zum Klartext. Dabei wird das Pseudonym durch Verschlüsselung des Klartextes generiert. „So kann bei Kenntnis des Verfahrens und des verwendeten Schlüssels das Pseudonym entschlüsselt und somit der zugrundeliegende Klartext aufgedeckt werden.“[71] Als eine weitere Verfügbarkeitsoption kann die Verkettbarkeit

[71] (Vgl. ebenda, S. 19)

von Pseudonymen betrachtet werden. Dabei wird geprüft, ob ein spezifischer Zusammenhang der zugrundeliegenden Klartexte der Pseudonyme gegeben ist.

Des Weiteren kann eine Verfügbarkeitsoption an bestimmte Zwecke oder Rollen gebunden sein. Dabei werden bestimmte Rollen oder Zwecke bestimmten Verfügbarkeitsoptionen zugeordnet. Im Folgenden werden Pseudonymisierungsverfahren mit spezifischen Verfügbarkeitsoptionen beschrieben.

Verkettbare aufdeckbare Pseudonyme:

Bei diesem Verfahren werden deterministische Verschlüsselungsverfahren angewandt. Dabei werden gleiche Klartexte auf gleiche Pseudonyme abgebildet. Somit existiert auch ohne Kenntnis des Schlüssels eine Verkettbarkeit der Pseudonyme. Eine Aufdeckbarkeit kann jedoch erst bei Kenntnis des Schlüssels gewährleistet werden.

Nicht-verkettbare aufdeckbare Pseudonyme:

Hierbei kommen probabilistische Verschlüsselungsverfahren zur Erstellung von nicht-verkettbaren aufdeckbaren Pseudonymen zum Einsatz. Diese Verschlüsselungsverfahren ermöglichen eine Abbildung von gleichen Klartexten auf unterschiedliche Pseudonyme. Somit kann keine Verkettbarkeit der Pseudonyme angenommen werden. Ist jedoch der Entschlüsselungsschlüssel bekannt, können die Pseudonyme aufgedeckt werden.

Verkettbare nicht-aufdeckbare Pseudonyme:

Verkettbare nicht-aufdeckbare Pseudonyme werden durch deterministische Einwegfunktionen, wie Hash-Funktionen, erstellt. Dabei werden gleiche Klartexte auf gleiche Pseudonyme abgebildet, sodass eine Verkettbarkeit zwischen den Pseudonymen gewährleistet ist. Da hierbei eine Einwegfunktion zum Einsatz kommt, kann die Umkehrung der Pseudonymisierung, also die Aufdeckung der Pseudonyme, ausgeschlossen werden.[72]

Deterministische Pseudonyme:

Hierbei werden die Pseudonyme durch schlüsselabhängige Einweg- oder Hashfunktionen von einer vertrauenswürdigen zentralen Instanz aus invarianten Daten, zum Beispiel Identitätsdaten, erzeugt.

Willkürliche Pseudonyme:

Der Benutzer erzeugt sein Pseudonym durch einen festen Einweg-Algorithmus, der zum Beispiel aus einem Geheimnis oder einer Passphrase entsteht.

Zufällige Pseudonyme:

Hierbei werden die Pseudonyme frei gewählt oder durch ein Zufallsverfahren erzeugt. Diese zufällig erzeugten Pseudonyme (Einmalpseudonyme) finden ihre Anwendung oft bei Forschungszwecken, wo es zu Zusammenführungen von unterschiedlichen Datenquellen kommen kann. Dabei sind sie nur wiederverwendbar, wenn sie zum Beispiel in einer Referenzliste gespeichert werden.[73]

---

72 (Vgl. ebenda, S. 20 f)

73 (Vgl. Pommerening 2005, S. 1)

Rollenbindung

Hierbei kann der Zugriff auf das ggf. verkettete oder aufdeckbare Pseudonym nur durch eine definierte Rolle erfolgen. Dabei wird der zur Aufdeckung erforderliche Entschlüsselungsschlüssel nur einer definierten Rolle zugeordnet.[74]

Hash-Funktion

Im Allgemeinen berechnen Hashfunktionen eine Art Prüfsumme (Hashwerte) für bestimmte Dateien. Dabei dient der Vergleich von Hashwerten beispielsweise der Kontrolle, ob Daten korrekt kopiert wurden.[75]

Abbildung 4-5 soll die Definition verdeutlichen:

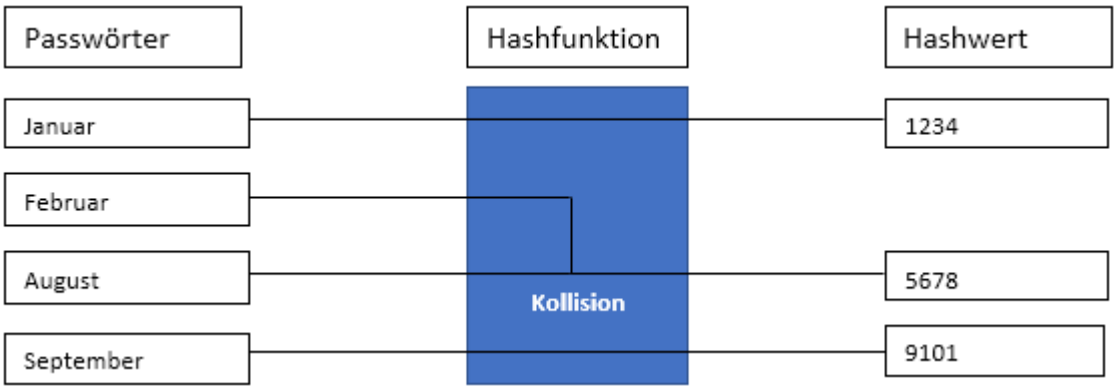

Abbildung 4-5: Veranschaulichung einer Kollision in einer Hash-Funktion

In der Abbildung befinden sich auf der linken Seite vier verschiedene Passwörter von Mitarbeitern in einem Unternehmen. Die Hashfunktion (math. Algorithmus), die sich in der Mitte befindet, wandelt diese Passwörter in eine Zeichenfolge, den Hashwert, um. Dieser besitzt eine bestimmte Länge, oft wird dabei eine hexadezimale Zeichenkette verwendet. Im Beispiel besteht der Hashwert aus 4 Zeichen.

Anschließend wird jedem Passwort ein Hashwert zugeordnet. „Januar“ besitzt den Hashwert „1234“, „Februar“ und „August“ besitzen den Hashwert „5678“ und „September“ den Hashwert „9101“.[76]

Zusammenfassend wandeln Hashfunktionen Zeichenfolgen unterschiedlicher Längen in Zeichenfolgen einer bestimmten Länge um.[77]

Die nachfolgenden Eigenschaften stellen die Anforderungen an eine Hashfunktion dar. Wird eine der Anforderungen nicht erfüllt, gilt eine Hashfunktion als gebrochen und sollte nicht mehr eingesetzt werden.

---

74 (Vgl. Schwartmann und Weiß 2017, S. 21)
75 (Vgl. Czernik 2016)
76 (Vgl. ebenda)
77 (Vgl. Kurose und Ross 2012, S. 744)

Einwegfunktion

Es darf nicht möglich sein, dass der Hashwert auf den Originalwert zurückzuführen ist.

Kollisionssicherheit

Hierbei darf den unterschiedlichen Zeichenfolgen nicht derselbe Hashwert zugeordnet werden. [78] Wird dieser Punkt erfüllt, ist eine kryptographische Hashfunktion vorhanden. Im oben aufgeführten Beispiel findet jedoch eine Kollision statt, da den Passwörtern „Februar“ und „August“ derselbe Hashwert zugeordnet wurde. Diese Hashfunktion ist somit nicht kollisionssicher und somit auch keine kryptographische Hashfunktion. [79]

Schnelligkeit

Es muss sichergestellt werden, dass die Berechnung des Hashwertes schnell und ohne Verzögerung verläuft.[80]

4.3 ANONYMISIERUNG

Anonymisieren war bis zum 24.05.2018, vor dem Inkrafttreten der EU DS-GVO, in § 3 Abs. 6 BDSG-alt als das "Verändern personenbezogener Daten in einer solchen Weise, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können", legaldefiniert.

Somit sind anonyme Daten das Gegenteil von personenbezogenen Daten und fallen nicht in den Anwendungsbereich des Datenschutzrechts. Dabei sind nicht nur solche Daten anonym, die sich von vornherein nicht auf eine identifizierte oder identifizierbare Person beziehen. Vielmehr handelt es sich auch um anonyme Daten, wenn personenbezogene Daten vom Verantwortlichen anonymisiert worden sind. Dies ergibt sich nun aus den Erwägungsgründen 26 S. 5 und 6 EU DS-GVO, wo es heißt:

„Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, d.h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann. Diese Verordnung betrifft somit nicht die Verarbeitung anonymer Daten, auch nicht für statistische oder für Forschungszwecke.”[81]

Die in Erwägungsgrund 26 Satz 5 und 6 angesprochenen Vorgaben sind nur schwer mit dem Verständnis des Datenschutzes zu vereinbaren, da durch die heutigen technischen Mittel durchaus die Möglichkeit besteht, aus anonymisierten Daten einen Personenbezug wiederherzustellen.

Die Anonymisierung bezeichnet Verfahren, durch die der Bezug der Daten zu natürlichen Personen so entfernt worden ist, dass dieser Bezug nicht mehr hergestellt werden kann oder zumindest die Daten nicht mehr die Definition der Personenbezogenheit nach Art. 4 Nr. 1 EU DS-GVO erfüllen.

---

78 (Vgl. Burgard und Stachniss 2008, S. 9)
79 (Vgl. Czernik 2016)
80 (Vgl. ebenda)
81 (Vgl. Erwägungsgrund 26 DSGVO)

Abbildung 4-6 zeigt beispielhaft eine anonymisierte Tabelle. Hier wurde durch die Aufnahme eines PLZ-Gebiets, einer Altersgruppe, einer Berufsgruppe und durch das Weglassen der Geschlechtsinformationen jegliche Bezugsmöglichkeit zu einer Person entfernt. In diesem Zusammenhang sei darauf hingewiesen, dass – wie später in Kapitel 5.4.1 beschrieben – die Anonymisierung on demand erfolgt und so u.U. das Geschlecht je nach Forschungsfrage in die anonymisierten Informationen mit aufgenommen oder weggelassen werden kann. Sollte die Anzahl der Datensätze relativ klein sein, kann beispielsweise eine Verallgemeinerung des PLZ-Bereichs Abhilfe schaffen.

PersAuswNr	Vorname	Nachname	Straße	PLZ	Ort	Geburtsdatum	Beruf	Geschlecht	Kredit
918273645	Max	Muster	Holunderbusch 3a	98765	Musterort	01.01.1960	Arzthelfer	m	20000

anonymisierte Tabelle

PLZ-Bereich	Altersgruppe	Berufsfeld	Kredit
98*	50 bis 59	Medizin	20000

Abbildung 4-6: anonymisierte Tabelle

4.3.1 ABSOLUTE VS. FAKTISCHE ANONYMITÄT

4.3.1.1 ABSOLUTE ANONYMITÄT

Die absolute Anonymisierung beschreibt, dass personenbezogene Daten durch Vergröberung oder durch Entfernung einzelner Merkmale so verändert werden, dass eine Identifizierung der ursprünglichen Daten unmöglich ist.[82]

Hier ist die Zuordnung einer Einzelangabe zu einer konkreten Person unmöglich. Dies geschieht durch eine Vergröberung der Daten oder durch eine Entfernung einzelner Merkmale in einer Weise, dass eine personenbezogene Zuordnung der ursprünglichen Daten unmöglich wird. Mit zunehmender Erhöhung der Leistung von Rechnern und Prozessoren sowie der Speicherkapazitäten und aufgrund einer immer leichteren Verknüpfbarkeit von Daten lässt sich aber immer seltener feststellen, dass Daten „absolut anonym" sind und sich jeglicher Zuordnung zu einer Person entziehen. Entsprechende Anonymisierungsverfahren, die zu einer „absoluten Anonymität" führen könnten, lassen sich kaum noch durchführen.

4.3.1.2 FAKTISCHE ANONYMITÄT

Die Zuordnung einer Einzelangabe zu einer konkreten Person ist hier zwar nicht ausgeschlossen, erfordert jedoch einen unverhältnismäßig großen Aufwand in Bezug auf Zeit, Kosten und Arbeitskraft.[83] Das übliche Ergebnis einer Anonymisierung ist diese „faktische Anonymität".

Die EU DS-GVO unterscheidet nun nicht mehr zwischen „absoluter" und „faktischer" Anonymität, da der Begriff der Anonymität nur noch in Erwägungsgrund 26 Satz 5 und 6 EU DS-GVO genannt wird.

Werden anonyme bzw. anonymisierte Daten zu statistischen Zwecken oder zu Forschungszwecken verarbeitet, lässt Erwägungsgrund 26 Satz 6 EU DS-GVO den Einwand zu, dass diese Daten einer Anwendung der EU DS-GVO entzogen sind.

---

82 (Vgl. Statistische Ämter des Bundes und der Länder Forschungsdatenzentren)

83 (Vgl. Plath und Schreiber in Plath 2018, Rn. 59)

Erwägungsgrund 26 Satz 5 EU DS-GVO lässt sich schwer mit einem „absoluten“ Verständnis des Personenbezugs vereinbaren, da sich unter den heutigen Bedingungen der Datentechnik aus den oben genannten Gründen eine Identifizierung nie sicher ausschließen lässt.

Es dürfte aber weiterhin die faktische Anonymität ausreichen, die vorliegt, wenn mit nach allgemeinem Ermessen wahrscheinlich zu nutzenden Mitteln kein Personenbezug wiederhergestellt werden kann.[84]

Im Grundsatz besteht die einfachste Methode der Anonymisierung darin, direkte Identifikationsmerkmale wie den Namen des Betroffenen vom Datensatz zu entfernen oder unlesbar zu machen. Allerdings reicht dies für eine Anonymisierung nicht aus, wenn andere Umstände herangezogen werden können, durch die der Betroffene ausreichend klar individualisiert werden kann.

Für den Bereich Statistik wird nach § 12 Abs. 1 BStatG eine Anonymisierung herbeigeführt, indem zunächst die Hilfsmerkmale von den Erhebungsmerkmalen getrennt werden. Bei den Hilfsmerkmalen handelt es sich um die personenbezogenen Daten des Befragten, während nur die in den Erhebungsmerkmalen enthaltenen Informationen Gegenstand der Statistik sind. Die Erhebungsmerkmale aller Betroffenen werden sodann zumindest in solch übergeordneten Gruppen zusammengefasst, dass individuelle Rückschlüsse nicht mehr möglich sind. Weitere moderne Anonymisierungstechniken beruhen auf mathematischen und statistischen Verfahren und erzeugen z.B. Zufallsfehler in den Daten oder vertauschen einzelne Daten innerhalb einer Gruppe (Randomisierung).[85] Auf diese Weise sollen die entsprechenden Merkmale nicht mehr personenbezogenen Daten zugeordnet werden können, wobei zugleich der statistische Informationsgehalt erhalten bleiben soll. Um dies zu gewährleisten, muss vorab der De-Anonymisierungsaufwand mit Hilfe der unterschiedlichen Anonymisierungsverfahren (Aggregation, Klassenbildung von Merkmalsträgern usw.) analysiert werden.

Bei der faktischen Anonymisierung kann die De-Anonymisierung von Daten nicht komplett ausgeschlossen werden. Hierbei können die Daten jedoch nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft den jeweiligen Merkmalsträgern zugeordnet werden (§16 Abs. 6 BstatG). Das Ziel der faktischen Anonymisierung ist es durch die Minimierung und Veränderung von Informationen, die Zuordnungsmöglichkeiten von entsprechenden Merkmalen zu den personenbezogenen Daten zu verringern, dabei soll jedoch der statistische Informationsgehalt erhalten bleiben. Um dies zu gewährleisten, muss vorab der De-Anonymisierungsaufwand mit Hilfe der unterschiedlichen Anonymisierungsverfahren (Aggregation, Klassenbildung von Merkmalsträgern, usw.) analysiert werden.[86]

In Abbildung 4-7 ist exemplarisch dargestellt, wie aus unzureichend anonymisierten Tabellen durch Kombination mehrerer Datenquellen zusätzliche Informationen mit einer relativ hohen Wahrscheinlichkeit zugeordnet werden können. Im Beispiel liegt eine sehr hohe Überschneidungsmenge vor. Die in den unzureichend anonymisierten Tabellen vorliegenden Informationen zur Höhe des Kredits und zum Interesse des Individuums können mit einer hohen Wahrscheinlichkeit einer bestimmten Person zugeordnet werden. Je mehr Datenquellen zur Verfügung stehen, desto höher ist das entstehende Überschneidungswissen.

---

84 (Vgl. Schantz in Schantz/Wolff 2017, Kap. C Rn. 297); (Vgl. Art. 29-Gruppe, Stellungnahme 5/2014 zu Anonymisierungstechniken, WP 216 v. 10.04.2016)

85 (Vgl. ebenda)

86 (Vgl. Statistische Ämter des Bundes und der Länder Forschungsdatenzentren)

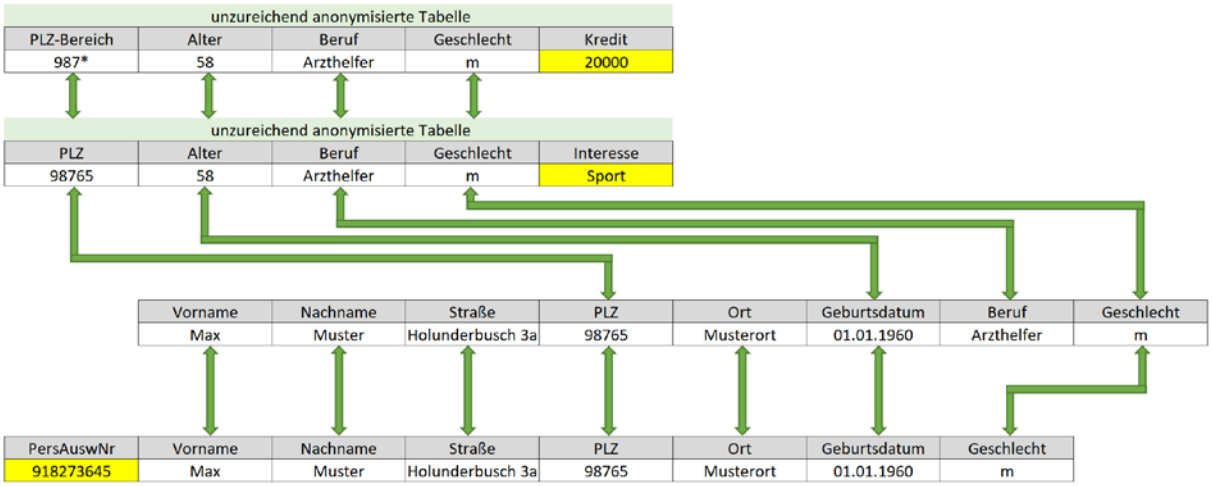

Abbildung 4-7: Generierung von Zusatzwissen bei unzureichend anonymisierten Daten

4.3.2 RECHTLICHE RISIKEN BEIM VERLUST DER ANONYMITÄT

Wenn eine Anonymisierung aufgehoben wird oder sonst verloren geht und die betroffene Person auf diese Weise wieder identifiziert werden kann, dann stellt dieser Vorgang eine besonders schwere Beeinträchtigung der Rechte des Betroffenen dar.

Die besondere Schwere der Rechtsverletzung ergibt sich einerseits daraus, dass der Betroffene gerade aufgrund der Schutzmaßnahmen der Anonymisierung nicht mit einer Identifizierung rechnen musste. Diese Erwartungshaltung wird auch in den Erwägungsgründen 47 und 50 der EU DS-GVO berücksichtigt. So überwiegen nach Erwägungsgrund 47 die Interessen und Grundrechte des Betroffenen am Schutz seiner Daten gegenüber dem Interesse des Verantwortlichen an der Datenverarbeitung insbesondere dann, wenn personenbezogene Daten in Situationen verarbeitet werden, in denen eine betroffene Person vernünftigerweise nicht mit einer weiteren Verarbeitung rechnen muss. Erwägungsgrund 50 nimmt an, dass die Weiterverarbeitung personenbezogener Daten für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke als vereinbarer und rechtmäßiger Verarbeitungsvorgang gilt, wenn es sich um den ursprünglichen Zweck handelt.

Andererseits wird in Art. 6 Abs. 4 lit. e) EU DS-GVO gerade die Pseudonymisierung oder Verschlüsselung als ein Merkmal genannt, das für eine Vereinbarkeit der Datenverarbeitung mit dem Zweck der ursprünglichen Datenerhebung spricht. Daraus folgt, dass ein Verlust der Pseudonymisierung, Anonymisierung oder Verschlüsselung zur Unvereinbarkeit der Datenverarbeitung mit dem Erhebungszweck führen dürfte.

Im Ergebnis führt eine Aufhebung der Anonymisierung also dazu, dass die Verarbeitung der Daten nicht mehr vom Verarbeitungszweck erfasst ist und damit unzulässig wird. Da es sich sodann um eindeutig personenbezogene Daten handelt, finden die EU DS-GVO und das BDSG sodann auch in vollem Umfang Anwendung. Aufgrund der Schwere der Rechtsverletzung der Betroffenen dürften somit alle Daten umgehend zu löschen sein.

4.3.3 ANFORDERUNGEN AN ANONYMITÄT

In diesem Kapitel werden die Anforderungen an die Anonymität definiert, damit eine De-Anonymisierung von personenbezogenen Daten verhindert werden kann.

Dazu gehören die Verhinderung von Identity Disclosure, Attribute Disclosure und Membership Disclosure. Weiterhin dürfen die Anonymisierungsverfahren keinen Einfluss auf die Datenqualität haben und es sollte die Schutzreserve, bezüglich des technischen Fortschritts mit einbezogen werden, um zukünftige Risiken einer De-Anonymisierung durch den technischen Fortschritt vorzubeugen. Im Folgenden werden diese Anforderungen noch einmal definiert.

Vermeidung von Identity Disclosure

Identity Disclosure beschreibt die Zuordnung eines Individuums zu einem bestimmten Datensatz. Somit hat der Angreifer die Möglichkeit, die sensiblen Informationen eines Individuums zu erfahren.[87] Diese Zuordnung zu einem Individuum soll mit einem geeigneten Anonymisierungsverfahren vermieden werden.

Vermeidung von Attribute Disclosure

Eine Vermeidung von Attribute Disclosure soll ebenfalls durch ein bestimmtes Anonymisierungsverfahren gewährleistet sein. Allgemein kann hierbei auch ohne eine direkte Zuordnung einer Person zu einem bestimmten Eintrag, ein Angreifer sensible Daten aufdecken. Es reicht aus, wenn die Möglichkeit besteht, den Wert eines Attributs mit einem Individuum zu verknüpfen.[88] Dies kann auftreten, wenn die vorhandenen Einträge mit den gleichen identifizierenden Attributen und den sensitiven Attributen übereinstimmen.[89] Beispielsweise könnte ein Angreifer mit einer sehr hohen Wahrscheinlichkeit davon ausgehen, dass eine bestimmte Person unter HIV leidet, wenn viele Merkmale mit Eigenschaften einer Person, die dem Angreifer bekannt ist, übereinstimmen.

Vermeidung von Membership Disclosure

Membership Disclosure tritt auf, wenn aufgedeckt werden kann, dass sich ein Individuum mit einer hohen Wahrscheinlichkeit in einem bestimmten Datensatz befindet. Diese Zuordnung eines Individuums zu einem bestimmten Datensatz kann zu Problemen führen und sollte somit durch eine geeignete Anonymisierung der personenbezogenen Daten verhindert werden. Das folgende Szenario führt dieses Problem genauer auf.

Wenn es möglich ist eine Zugehörigkeit einer bestimmten Person zu einer, für Forschungszwecke generierten Tabelle mit Informationen über HIV-Patienten zu ermitteln, kann durch diese reine Information der Zugehörigkeit, möglicherweise ein großer Schaden für ein Individuum entstehen.[90]

Gewährleistung der Datenqualität

Eine weitere Anforderung an das Anonymisierungsverfahren stellt die Gewährleistung der Datenqualität dar. Die Anonymisierung der personenbezogenen Daten sollte kaum bis keinen Einfluss auf die

---

87 (Vgl. Li und Li 2008, S. 446)
88 (Vgl. Bender 2015, S. 13)
89 (Vgl. Li und Li 2008, S. 446)
90 (Vgl. Bender 2015, S. 14) und (Vgl. Li und Li 2008, S. 446)

Qualität der Datensätze haben, sodass der Nutzen der Daten für verschiedene Auswertungen erhalten bleibt.

Einbeziehung der Schutzreserve

Um eine De-Anonymisierung von personenbezogenen Daten zu verhindern, sollte die „Schutzreserve“ mit einbezogen werden. Diese soll den schnellen technologischen Fortschritt mit einbeziehen, denn ein Verfahren, das gegenwärtig eine ausreichende Anonymität der Daten gewährleistet, kann gegebenenfalls in ein paar Jahren diesen Schutz nicht mehr garantieren.[91]

4.3.4 METHODEN ZUR ANONYMISIERUNG

Bezüglich der Anonymisierung von personenbezogenen Daten müssen Regeln aufgestellt werden, die garantieren, dass eine De-Anonymisierung nicht möglich ist. So ist beispielsweise eine „background knowledge attack“, also die Möglichkeit, mit Hilfe von korrelierendem Wissen eine De-Anonymisierung durchzuführen, in jedem Fall zu verhindern. In diesem Zusammenhang wird deutlich, dass eine Mindestanzahl an anonymisierten Datensätzen pro Datengruppe existieren muss. Die Informationen müssen in der Art anonymisiert werden, dass einzelne Merkmale nicht dazu führen, dass einzelne Personen oder Personengruppen ausgeschlossen werden können und so die Anzahl der Datensätze zu klein wird.

Ein sehr einfaches, aber dennoch effektives Verfahren zur Anonymisierung ist die Generalisierung von Informationen, d.h. Informationen werden vergröbert (Beispiel: Akne oder Schuppenflechte werden zu Hautkrankheit verallgemeinert). Im Allgemeinen unterscheidet man zwischen lokalem Recordieren und globalem Recordieren.[92]

Beim lokalen Recordieren können für dieselben Werte innerhalb eines Attributs unterschiedliche Generalisierungen angewendet werden, während bei globalem Recordieren immer dieselbe Generalisierung für dieselben Werte angewendet wird. Kennzeichnend für das globale Recordieren ist neben der auf Grund des restriktiveren Modells etwas eingeschränkten Möglichkeiten bzgl. der Generalisierung die bessere Beherrschbarkeit des Suchraums.[93]

Weiterhin ist eine Unterscheidung in eindimensionale und mehrdimensionale Generalisierung möglich. Bei der eindimensionalen Generalisierung wird jedes Attribut unabhängig von weiteren Attributen generalisiert, während bei der mehrdimensionalen Generalisierung unterschiedliche Kombinationen und Generalisierungswerte erzeugt werden. So könnten beispielsweise die Attributwerte „12345“ (PLZ) und „Akne“ (Erkrankung) zu „123**“ und „Akne“ oder aber auch zu „12345“ und „Hautkrankheit“ oder zu „123**“ und „Hautkrankheit“ generalisiert werden.[94]

---

91 (Vgl. Schaar 2016, S. 8)
92 (Vgl. Fung et al. 2011, S. 376)
93 (Vgl. Kohlmayer 2015, S. 21)
94 (Vgl. Kohlmayer 2015, S. 21)

Im Beispiel in Tabelle 4-1 werden die Daten in gleich große Altersintervalle von jeweils fünf Jahren eingeteilt.

Tabelle 4-1: Daten mit gleich großen Altersintervallen

Geschlecht	Altersgruppe	PLZ	Augenfarbe	Diagnose	Raucher
m	45 – 49	46	blau	Herzinfarkt	ja
w	45 – 49	44	braun	Krebs	nein
m	110 – 114	58	grün	Krebs	ja
m	80 – 84	92	blau	Herzinfarkt	ja
w	45 – 49	34	braun	HIV	ja
w	45 – 49	56	braun	Herzinfarkt	nein
w	55 – 59	65	blau	HIV	ja
m	80 – 84	24	grün	Herzinfarkt	nein
w	30 – 34	33	blau	HIV	nein
m	55 – 59	80	blau	Krebs	nein
w	100 – 104	73	braun	Krebs	ja
w	90 – 94	92	braun	Herzinfarkt	ja
m	45 – 49	28	blau	HIV	nein
m	90 – 94	19	blau	Herzinfarkt	nein
w	55 – 59	06	grün	Krebs	ja
m	100 – 104	28	braun	Herzinfarkt	ja
m	90 – 94	33	blau	Krebs	nein
w	55 – 59	87	braun	Herzinfarkt	nein
w	90 – 94	56	grün	Krebs	ja
w	45 – 49	44	blau	HIV	nein

Ist die Information, welche Krankheit eine Person hat oder für welche Produkte sich jemand interessiert, in der Altersgruppe 40 bis 45 Jahre auf Grund der Gruppengröße noch anonym, so wird die Gruppengröße mit zunehmendem Alter immer geringer, bis schließlich einzelne zusätzliche Informationen wie Geschlecht, Wohnort oder ähnliches dazu führen können, dass Informationen einer einzelnen Person direkt zuzuordnen sind. So existiert im Beispiel in Tabelle 4-1 lediglich ein Datensatz in der Altersgruppe 110 - 114.

Aus diesem Grund müssen Intervalle, die zur Anonymisierung beitragen, immer so gewählt werden, dass eine genügend große Anzahl von Datensätzen pro Gruppe existiert. Ein Intervall für eine

Altersgruppe von 100 bis 109 und ein Intervall von 110 bis 119 Jahre wird immer relativ wenige Datensätze generieren. Empfehlenswert ist in einem solchen Fall eine größere Gruppe, beispielsweise eine Altersgruppe von Personen über 90 Jahre. Auch eine Granulation der PLZ-Gebiete führt zu einer detaillierteren Zuordnungsmöglichkeit und ist somit nicht ratsam.

Tabelle 4-2 zeigt, dass durch die Zuordnung der älteren Menschen zu einer Altersgruppe >90 eine ausreichend große Anzahl von Datensätzen für diese Gruppe existiert.

Tabelle 4-2: Daten mit unterschiedlich großen Altersintervallen

Geschlecht	Altersgruppe	PLZ	Augenfarbe	Diagnose	Raucher
m	45 – 49	46	blau	Herzinfarkt	ja
w	45 – 49	44	braun	Krebs	nein
m	>90	58	grün	Krebs	ja
m	80 – 90	92	blau	Herzinfarkt	ja
w	45 – 49	34	braun	HIV	ja
w	45 – 49	56	braun	Herzinfarkt	nein
w	55 – 59	65	blau	HIV	ja
m	80 – 90	24	grün	Herzinfarkt	nein
w	30 – 34	33	blau	HIV	nein
m	55 – 59	80	blau	Krebs	nein
w	>90	73	braun	Krebs	ja
w	>90	92	braun	Herzinfarkt	ja
m	45 – 49	28	blau	HIV	nein
m	>90	19	blau	Herzinfarkt	nein
w	55 – 59	06	grün	Krebs	ja
m	>90	28	braun	Herzinfarkt	ja
m	>90	33	blau	Krebs	nein
w	55 – 59	87	braun	Herzinfarkt	nein
w	>90	56	grün	Krebs	ja
w	45 – 49	44	blau	HIV	nein

Eine weitere Möglichkeit dieses Problem zu umgehen, besteht darin Gruppen, die nur sehr wenige Daten-sätze aufweisen, nicht zu übertragen. In diesem Fall muss aber geprüft werden, ob die Forschungsfrage dann noch korrekt beantwortet werden kann.

Im Folgenden sollen einige Arten zur Anonymisierung von Daten vorgestellt werden.

4.3.4.1 K-ANONYMITÄT

Beschreibung

Wenn personenbezogene Daten anonymisiert werden sollen, werden zuerst alle Identifikatoren wie Name, PLZ oder ähnliche Angaben, die zu einer leichten Identifikation führen können, gelöscht. Findet nach dem Streichen von leichten Identifikatoren eine Kombination der anonymisierten Daten mit externen Datenblöcken statt, kann es zu einer De-Anonymisierung kommen. Hierbei kann eine Abgleichung mit den Quasi-Identifiern wie Postleitzahlen, Geburtstagen, Geschlechter oder Familienständen mit Wählerverzeichnissen oder sozialen Netzwerken stattfinden.[95]

Im Detail funktioniert die k-Anonymität folgendermaßen: k beschreibt eine Zahl, die Auskunft über die Stärke der Anonymisierung gibt. Je höher k ist, desto stärker ist die Anonymisierung. Als erstes müssen die Quasi-Identifier erkannt und verallgemeinert werden. Dabei wird zum Beispiel der Geburtsort durch den Landkreis oder das Alter durch eine entsprechende Altersgruppe (z.B. 20-25-Jährige) ersetzt. Die Daten werden dann solange verallgemeinert, bis es zu jedem Datensatz „k-1" Datenzwillinge gibt.[96] Zwar kommt es bei der Verallgemeinerung der Daten zu einem minimalen Informationsverlust, aber bei einer Gefährdung kann kein einzelner Datensatz mehr identifiziert werden, sondern nur noch eine Gruppe von k-Datensätzen.[97]

Im biomedizinischen Bereich wird häufig ein Wert von k=5 verwendet, wobei ein Wertebereich von k=3 bis k=25 als normal angesehen wird.[98]

---

95 (Vgl. Gatzke 2012, S. 3)
96 (Vgl. Machanavajjhala et al. 2006, S. 6 ff.)
97 (Vgl. Interview mit Johann Eder 2018)
98 (Vgl. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine 2015, S. 280)

Beispiel: 3-Anonymität

Tabelle 4-3: k-Anonymität mit k=3

Quasi-Identifikator			sensibles Attribut
Geschlecht	Geburtsjahr	PLZ	Krankheit
M	1960	44141	Haarausfall
M	1960	44141	Akne
M	1960	44141	Heuschnupfen
M	1960	44141	Diabetes
W	1960	44141	Heuschnupfen
W	1960	44141	Akne
W	1960	44141	Erkältung
M	1961	44141	Erkältung
M	1961	44141	Heuschnupfen
M	1961	44141	Haarausfall
M	1961	44141	Akne
W	1961	44141	Haarausfall
W	1961	44141	Heuschnupfen
W	1961	44141	Erkältung
W	1961	44141	Akne
M	1962	44141	Heuschnupfen
M	1962	44141	Heuschnupfen
M	1962	44141	Heuschnupfen
M	1962	44141	Heuschnupfen

Tabelle 4-3 befindet sich in 3-Anonymität, da es mindestens drei Datensätze mit denselben Quasi-Identifikatoren gibt. Je höher die Werte von k sind, desto stärker ist die Anonymisierung.

Probleme

Die k-Anonymität bringt aber auch Probleme mit sich, die durch potentielle Angreifer ausgenutzt werden können. Zum einen kann es zu einem Homogenitätsangriff oder durch korrelierendes

Hintergrundwissen zu einer Zusammenführung von personenbezogenen Daten kommen.[99] Weitere Probleme im Zusammenhang mit k-Anonymität sind das Problem des "unsorted Matching", der Homogenitätsangriff sowie die Anwendung korrelierenden Wissens.

Unsorted Matching

Gibt man die Zeilen einer Tabelle aus, so werden Sie im Normalfall in einer Standardsortierung ausgegeben. Hier ist insofern ein Problem zu sehen, als dass bei einer Ausgabe von zwei oder mehr Teilergebnissen, die eigentlich nicht bei einem Dritten zusammengeführt werden sollen, eine Zuordnung der einzelnen Datensätze erfolgen kann, wenn dieser die Daten zufällig oder widerrechtlich erhält.

Beispiel: Komplette Ausgabe der Datensätze

Als Ausgangssituation in Tabelle 4-4 werden alle Datensätze in einer Tabelle ausgegeben. Eine direkte Zuordnung der Krankheit zu einer Person ist möglich.

Tabelle 4-4: Ausgangstabelle (Beispiel)

Vorname	Nachname	GebJahr	PLZ	Geschlecht	Krankheit
Hans	Meier	1960	44141	M	Haarausfall
Klaus	Berger	1960	44141	M	Akne
Klaus	Krause	1960	44141	M	Heuschnupfen
Adam	Himmel	1960	44141	M	Diabetes
Tanja	Hansen	1960	44141	W	Heuschnupfen
Frauke	Peters	1960	44141	W	Akne
Melanie	Maus	1960	44141	W	Erkältung
Harald	Meyer	1961	44141	M	Erkältung
Kurt	Spieß	1961	44141	M	Heuschnupfen
Thomas	Konrad	1961	44141	M	Haarausfall
Otto	Klein	1961	44141	M	Akne
Barbara	Schmidt	1961	44141	W	Haarausfall
Eva	Kunze	1961	44141	W	Heuschnupfen
Petra	Fritze	1961	44141	W	Erkältung
Erna	Müller	1961	44141	W	Akne

[99] (Vgl. LDI NRW 2017, S. 7)

Beispiel: Sortierte Ausgabe der Datensätze

Im Beispiel in Tabelle 4-5 ermitteln zwei Abfragen jeweils ein Ergebnis, dass ohne Kenntnis des jeweils anderen Ergebnisses keine Zuordnung der Krankheit zu einer Person ermöglicht. Allerdings kann dennoch eine Zuordnung einer Krankheitsinformation zu einer Person erfolgen, wenn beide Ergebnisse bei einem Dritten zusammengeführt werden, da die Datensätze jeweils in Standardsortierung ausgegeben wurden:

Tabelle 4-5: Ausgabe der Daten jeweils in Standardreihenfolge

Vorname	Nachname	GebJahr		PLZ	Geschlecht	Krankheit
Hans	Meier	1960	⟷	44141	M	Haarausfall
Klaus	Berger	1960	⟷	44141	M	Akne
Klaus	Krause	1960	⟷	44141	M	Heuschnupfen
Adam	Himmel	1960	⟷	44141	M	Diabetes
Tanja	Hansen	1960	⟷	44141	W	Heuschnupfen
Frauke	Peters	1960	⟷	44141	W	Akne
Melanie	Maus	1960	⟷	44141	W	Erkältung
Harald	Meyer	1961	⟷	44141	M	Erkältung
Kurt	Spieß	1961	⟷	44141	M	Heuschnupfen
Thomas	Konrad	1961	⟷	44141	M	Haarausfall
Otto	Klein	1961	⟷	44141	M	Akne
Barbara	Schmidt	1961	⟷	44141	W	Haarausfall
Eva	Kunze	1961	⟷	44141	W	Heuschnupfen
Petra	Fritze	1961	⟷	44141	W	Erkältung
Erna	Müller	1961	⟷	44141	W	Akne

Beispiel: Zufällige Ausgabe der Datensätze

Durch die Sortierung der Zeilen nach Zufall bei beiden Teilergebnissen im Beispiel in Tabelle 4-6 ist eine Zuordnung der sensiblen Krankheitsinformation nicht mehr möglich, auch dann nicht, wenn beide Ergebnisse bei einem Dritten zusammengeführt werden.

Tabelle 4-6: Ausgabe der Daten in zufälliger Reihenfolge

Vorname	Nachname	GebJahr
Harald	Meyer	1961
Melanie	Maus	1960
Tanja	Hansen	1960
Erna	Müller	1961
Thomas	Konrad	1961
Frauke	Peters	1960
Barbara	Schmidt	1961
Otto	Klein	1961
Kurt	Spieß	1961
Klaus	Krause	1960
Klaus	Berger	1960
Adam	Himmel	1960
Hans	Meier	1960
Petra	Fritze	1961
Eva	Kunze	1961

PLZ	Geschlecht	Krankheit
44141	M	Akne
44141	M	Haarausfall
44141	M	Heuschnupfen
44141	W	Heuschnupfen
44141	W	Akne
44141	M	Diabetes
44141	M	Haarausfall
44141	W	Haarausfall
44141	W	Akne
44141	W	Erkältung
44141	M	Akne
44141	W	Heuschnupfen
44141	W	Erkältung
44141	M	Erkältung
44141	M	Heuschnupfen

Homogenitätsangriff

k-Anonymität kann dazu führen, dass durch die Verallgemeinerung der Gruppen nicht nur die Quasi-Identifier die gleichen Werte besitzen, sondern auch die zu schützenden personenbezogenen Werte gleich sind.[100]

[100] (Vgl. LDI NRW 2017, S. 7)

Tabelle 4-7: 3-anonyme Tabelle mit gleichen Einträgen in den sensiblen Daten

Quasi-Identifikator			sensibles Attribut
Geschlecht	Geburtsjahr	PLZ	Krankheit
M	1960	44141	Haarausfall
M	1960	44141	Akne
M	1960	44141	Heuschnupfen
M	1960	44141	Diabetes
W	1960	44141	Heuschnupfen
W	1960	44141	Akne
W	1960	44141	Erkältung
M	1961	44141	Erkältung
M	1961	44141	Heuschnupfen
M	1961	44141	Haarausfall
M	1961	44141	Akne
W	1961	44141	Heuschnupfen
W	1961	44141	Heuschnupfen
W	1961	44141	Heuschnupfen
W	1961	44141	Heuschnupfen

In Tabelle 4-7 bilden die letzten vier Einträge eine Äquivalenzklasse, in der jeder Patient mit Heuschnupfen diagnostiziert werden kann. Somit kann jeder Patient aus der PLZ 44141 zwar keinem konkreten Eintrag zugeordnet werden, sein Befund ist jedoch erkennbar. Ein Angreifer kann daher mit dem Hintergrundwissen, dass eine Nachbarin im Jahr 1961 geboren wurde, aus der PLZ 44141 kommt und Patientin ist, sicher sein, dass diese Nachbarin unter Heuschnupfen leidet.

Bei der k-Anonymität wird somit eine konkrete Zuordnung von einem Individuum zu einem bestimmten Eintrag ausgeschlossen, jedoch kann ein Angreifer sensible Informationen enthüllen. Dies folgt daraus, dass alle Einträge mit denselben Quasi-Identifiern auch dieselben sensitiven Attribute besitzen.

Korrelierendes Wissen

Ein korrelierendes Wissen stellt bei der k-Anonymität ein großes Problem dar, es kann hierbei schnell durch Ausschlussverfahren zu einer Offenlegung von personenbezogenen Daten kommen. Somit

kann bei einem großen Hintergrundwissen keine vollständige Anonymisierung gewährleistet werden.[101]

Im Beispiel in Tabelle 4-7 könnte eine Person, die einen Patienten kennt, der 1960 geboren wurde und aus der PLZ 44141 stammt ableiten, dass dieser Patient an Diabetes leidet, wenn er bei einem zufälligen Treffen sieht, dass der Bekannte weder an Akne oder Haarausfall leidet, noch Symptome für Heuschnupfen zeigt.

4.3.4.2 L-DIVERSITY

Beschreibung

l-Diversity ist eine Weiterentwicklung der k-Anonymität, hierbei werden sensible und nicht-sensible Attribute unterschieden. Die sensiblen Attribute zählen dabei nicht zu den Quasi-Identifiern. Es werden lediglich identische Attributwerte des Quasi-Identifiers betrachtet. l-Diversity kann angenommen werden, wenn in jedem Datenblock mindestens „l“ verschiedene Werte des sensiblen Attributs vorhanden sind.[102] Je höher dabei das „l“ gewählt wird, desto höher ist der Schutz der sensiblen Attribute.[103]

Bei l-Diversity soll also sichergestellt werden, dass die sensitiven Attribute in einer Äquivalenzklasse - im Gegensatz zur k-Anonymität - genügend unterschiedliche Werte annehmen können.

Tabelle 4-8 verdeutlicht den Zusammenhang: Hier sind in jeder Äquivalenzklasse mindestens drei verschiedene sensible Attribute vorhanden. Wenn bei einem Angriff ein Angreifer mit speziellem Hintergrundwissen hinsichtlich einer bestimmten Person in einer bestimmten Äquivalenzklasse einen sensitiven Wert ausschließen möchte, bestehen jedoch in jeder Klasse noch mindestens zwei andere sensible Werte, die in Betracht gezogen werden können. Somit kann der Angreifer bezogen auf das Beispiel aus Kapitel 4.3.4.1 nicht ermitteln, welche Krankheit eine im Jahr 1961 geborene Frau aus der PLZ 44141 hat.

---

101 (Vgl. LDI NRW 2017, S. 7)
102 (Vgl. Petrlic und Sorge 2017, S. 37)
103 (Vgl. Hauf o.J., S. 9)

Tabelle 4-8: l-Diversity mit l=3

Quasi-Identifikator			sensibles Attribut
Geschlecht	Geburtsjahr	PLZ	Krankheit
M	1960	44141	Haarausfall
M	1960	44141	Akne
M	1960	44141	Heuschnupfen
M	1960	44141	Diabetes
W	1960	44141	Heuschnupfen
W	1960	44141	Akne
W	1960	44141	Erkältung
M	1961	44141	Erkältung
M	1961	44141	Akne
M	1961	44141	Diabetes
M	1961	44141	Diabetes
W	1961	44141	Heuschnupfen
W	1961	44141	Diabetes
W	1961	44141	Akne
W	1961	44141	Haarausfall

Probleme

Skewness Attack

Skewness-Attacken können dann angewendet werden, wenn die Gesamtverteilung der vorhandenen Attributwerte ungleichmäßig ist oder ein bestimmter Wert nur sehr selten auftritt. Als Beispiel kann eine Tabelle mit Patientendaten dienen. In dieser Tabelle gibt das sensitive Attribut an, ob bei einem Patienten HIV diagnostiziert wurde. Wenn dabei auch nur 1% aller Patienten mit HIV-positiv diagnostiziert wurden und in einer Äquivalenzklasse genauso viele Werte negativ wie positiv sind, erhöht sich die Wahrscheinlichkeit für jeden Patienten auf 50% an HIV erkrankt zu sein. Diese Informationen würden bereits ausreichen, um eine betroffene Person zu schaden.[104]

[104] (Vgl. Bender 2015, S. 14)

Similarity Attack

Die Similarity Attack, auch Ähnlichkeitsattacke genannt, kann stattfinden, wenn die Gruppe, in der ein Individuum anonymisiert wurde, nur wenige semantische Unterschiede aufweist.[105]

„Wenn zum Beispiel in einer medizinischen Tabelle, bei einer bestimmten Gruppe von Patienten, das sensitive Attribut als Wert ausschließlich verschiedene Arten von Krebs angibt, ist die Schlussfolgerung daraus, dass jeder Patient in dieser Gruppe an Krebs erkrankt ist."[106]

4.3.4.3 T-CLOSENESS

Beschreibung

Da l-Diversity nicht ausreichend vor einer Zuordnung einer Person zu einem bestimmten Eintrag in einem Datensatz schützen kann, soll im weiteren Verlauf das Verfahren t-Closeness betrachtet werden. Das Konzept t-Closeness ist eine Weiterentwicklung der k-Anonymität und der l-Diversity. Es bezieht zusätzlich die Semantik der personenbezogenen Daten ein. Das Ziel von t-Closeness besteht darin, den möglichen Datengewinn eines Angreifers in Bezug auf die sensitiven Attribute in einer Äquivalenzklasse zu reduzieren.

Es muss gewährleistet sein, dass sich die Verteilungen der sensitiven Attribute einer Äquivalenzklasse von den Gesamtverteilungen in einer Tabelle möglichst gering unterscheiden. Dazu wird der Schwellwert *t* berechnet.

Ist *t* dabei kleiner als die Distanz zwischen der Verteilung eines sensitiven Attributs in der einzelnen Klasse und der Verteilung des Attributs in der gesamten Tabelle, ist t-Closeness gegeben. Es gilt je kleiner *t*, desto höher ist die Anonymisierung.[107]

Durch das Verfahren sollen Informationen über den Zusammenhang von Quasi-Identifiern und sensitiven Attributen reduziert werden. Dies soll einerseits vor einer Zuordnung von einer bestimmten Person zu einem bestimmten Eintrag in einem Datensatz schützen und andererseits den Nutzen der vorhandenen Daten verringern. Der Parameter *t* beschreibt zusätzlich das Verhältnis von Nutzen und Sicherheit.[108]

Bestimmung der t-Closeness mit Hilfe der Kullback-Leibler-Divergenz

Die Kullback-Leibler-Divergenz ist wie folgt definiert:

$$KL(T||T') = \sum_{i=1}^{N} T_i * log_2 \frac{T_i}{T_i'}$$

Dabei bilden $p = (p_1, ..., p_n)$ und $q = (q_1, ..., q_n)$ die Wahrscheinlichkeitsverteilungen.

In einem zweiten Schritt werden sogenannte q*-Blöcke gebildet, für diese werden dann im folgenden Beispiel die Werte berechnet.

---

105 (Vgl. Dölle 2015, S. 49)
106 (Vgl. Bender 2015, S. 15)
107 (Vgl. Hauf o. J., S. 12 ff.)
108 (Vgl. Bender 2015, S. 13) und (Vgl. Li et al. 2007, S. 110)

Die Tabelle 4-9 enthält vier q*-Blöcke mit jeweils unterschiedlich vielen Tupeln. Dabei wird eine k-Anonymität von k=5 erreicht. Weiterhin wird eine l-Diversity von l=3 erfüllt, da in jedem sensitiven Attribut mindestens drei unterschiedliche Krankheiten vorkommen.[109]
Um die t-Closeness zu berechnen, werden die Wahrscheinlichkeiten für das Auftreten der einzelnen bestimmten Werte des sensitiven Attributs in Bezug auf die gesamte Tabelle und in Bezug auf die einzelnen q*-Blöcke betrachtet.

Wichtig dabei ist es, dass jeder Wert für jeden q*-Block berechnet werden muss, um schließlich den Wert der t-Closeness zu erhalten.[110] Hierbei gilt: je niedriger der Wert der t-Closeness ist, desto besser ist die Anonymisierung der Daten.

[109] (Vgl. Goltz 2017, S. 12 ff.)
[110] (Vgl. ebenda) und (Vgl. Li et al. 2007, S. 109 f.)

Tabelle 4-9: Tabelle (t-Closeness = 0,31242)

Quasi-Identifikator			sensibles Attribut
Geschlecht	Geburtsjahr	PLZ	Krankheit
M	1960	44141	Heuschnupfen
M	1960	44141	Akne
M	1960	44141	Heuschnupfen
M	1960	44141	Diabetes
M	1960	44141	Akne
M	1960	44141	Heuschnupfen
M	1960	44141	Diabetes
M	1960	44141	Diabetes
W	1960	44141	Heuschnupfen
W	1960	44141	Akne
W	1960	44141	Diabetes
W	1960	44141	Akne
W	1960	44141	Akne
M	1961	44141	Diabetes
M	1961	44141	Akne
M	1961	44141	Diabetes
M	1961	44141	Diabetes
M	1961	44141	Akne
M	1961	44141	Heuschnupfen
M	1961	44141	Heuschnupfen
M	1961	44141	Diabetes
W	1961	44141	Heuschnupfen
W	1961	44141	Diabetes
W	1961	44141	Diabetes
W	1961	44141	Diabetes
W	1961	44141	Diabetes
W	1961	44141	Akne
W	1961	44141	Heuschnupfen

Tabelle 4-10: Verteilung der Krankheiten pro Block

	Akne	Heuschnupfen	Diabetes
pT gesamte Verteilung	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{3}{7}$
PT‘ erster q*-Block	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$
PT“ Zweiter q*-Block	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
PT‘‘‘ Dritter q*-Block	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
PT‘‘‘‘ Vierter q*-Block	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{4}{7}$

$$KL(T||T') = \sum_{i=1}^{3} T_i * log_2 \frac{T_i}{T_i'}$$

$$KL(T||T') = \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{1}{4}} + \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{3}{8}} + \frac{3}{7} * log_2 \frac{\frac{3}{7}}{\frac{3}{8}}$$

$$KL(T||T') \approx 0{,}08997$$

$$KL(T||T'') = \sum_{i=1}^{3} T_i * log_2 \frac{T_i}{T_i''}$$

$$KL(T||T') = \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{3}{5}} + \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{1}{5}} + \frac{3}{7} * log_2 \frac{\frac{3}{7}}{\frac{1}{5}}$$

$$KL(T||T') \approx 0{,}02551$$

$$KL(T||T''') = \sum_{i=1}^{3} T_i * log_2 \frac{T_i}{T_i'''}$$

$$KL(T||T') = \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{1}{4}} + \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{1}{4}} + \frac{3}{7} * log_2 \frac{\frac{3}{7}}{\frac{1}{2}}$$

$$KL(T||T') \approx 0{,}31242$$

$$KL(T||T'''') = \sum_{i=1}^{3} T_i * log_2 \frac{T_i}{T_i''''}$$

$$KL(T||T') = \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{1}{7}} + \frac{2}{7} * log_2 \frac{\frac{2}{7}}{\frac{2}{7}} + \frac{3}{7} * log_2 \frac{\frac{3}{7}}{\frac{4}{7}}$$

$$KL(T||T') \approx 0{,}01477$$

Der Wert für die t-Closeness beträgt im Beispiel 0,31242, da

max(KL(T||T'), KL(T||T''), KL(T||T'''), KL(T||T'''')) = KL(T||T''') ≈ 0,31242 ist.

Probleme

In der nennenswerten Literatur sind aktuell keine Angriffe oder Einschränkungen bei der Verarbeitung dokumentiert.

4.3.4.4 DIFFERENTIAL PRIVACY

Beschreibung

Differential Privacy ist ein Lösungsvorschlag, um eine De-Anonymisierung von Daten mit Hilfe von Hintergrundinformationen zu verhindern.[111] Der Grund für den Einsatz von Differential Privacy liegt darin, dass durch sogenannte „linkage attacks" anonymisierte Daten mit nicht anonymisierten Daten kombiniert werden können, wodurch es letztendlich möglich ist, die Personen hinter den anonymisierten Daten eindeutig identifizieren zu können.[112]

Für Differential Privacy wird vorausgesetzt, dass der Anonymisierungsprozess randomisiert ist und keine Änderung an einem einzelnen Datensatz der ursprünglichen Daten einen wesentlichen Einfluss auf die entsprechende Ausgabenverteilung hat. Dadurch wird sichergestellt, dass einzelne Daten

111 (Vgl. Dwork und Roth 2014, S. 5)
112 (Vgl. Sweeney 2002, S. 2)

eines Individuums nicht im Anonymisierungsergebnis wiederzuerkennen sind und es auch keine Zusammenführung von Datensätzen durch spezielles Hintergrundwissen gibt.[113]

Zum besseren Verständnis von Differential Privacy kann ein Beispiel von Cynthia Dwork angeführt werden:[114]

Dwork bezieht sich auf die Gewährleistung der Privatheit mit Hilfe eines Zufallsfaktors. In diesem Versuch wird den Teilnehmern eine persönliche Frage gestellt. Bevor sie antworten, soll eine Münze geworfen werden. Bei Zahl antworten die Teilnehmer wahrheitsgemäß auf die gestellte Frage und bei Kopf wird die Münze erneut geworfen. Anschließend sollen die Teilnehmer bei Kopf mit „Ja" und bei Zahl mit „Nein" antworten. Am Ende kann nun nicht mehr sicher gesagt werden, ob das Ergebnis wahr ist oder durch einen Zufall verfälscht wurde. Aus dem Gesamtergebnis können, durch Herausrechnen des Rauschfaktors, wenn dieser bekannt ist, jedoch weiterhin brauchbare Aussagen gemacht werden.

Im Folgenden wird gezeigt, wie Differential Privacy mathematisch definiert ist und welche Mechanismen dazu verwendet werden.

Mathematisches Modell:

ε-Differential-Privacy ist definiert als:[115]

$$\mathbb{P}(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times \mathbb{P}(\kappa(D_2) \in S)$$

Mit

D_1, D_2: benachbarte Datensätze, die sich in genau einem Eintrag unterscheiden

κ: Zufallsmechanismus

S: Bildmenge des Zufallsmechanismus

ε: Parameter, dieser bestimmt, wie stark sich das Fehlen eines einzelnen Datensatzes auf das Abfrageergebnis auswirkt.

Hash-Verfahren

Wie schon in Kapitel 4.2 beschrieben, werden beim Hashing personenbezogene Daten in eine zufällig angeordnete Zeichenfolge, die nicht mehr in ihren ursprünglichen Zustand zurückgeführt werden kann, überführt.[116]

Unter verschiedenen Umständen kann diese Art der Anonymisierung jedoch durch einen Abgleich mit mehreren Datenbanken aufgedeckt werden. Deshalb können Verschlüsselungen mittels Hash-Verfahren als ein Baustein der Differential Privacy, jedoch nicht als ein alleiniges Verfahren, betrachtet werden.

Subsampling-Verfahren

Beim Subsampling wird nur ein Teil der vorhandenen Datenzeilen zufällig ausgewählt und freigegeben. Die gewünschten Statistiken können mit Hilfe der Teilstichproben berechnet werden. Wenn

---

113 (Vgl. Bode et al. 2017, S. 23)
114 (Vgl. Dwork und Roth 2014, S. 30 ff.)
115 (Vgl. ebenda)
116 (Vgl. Els 2017, S. 221)

diese ausreichend groß sind, können sie repräsentativ für den ganzen Datensatz angesehen werden. Ist die Größe der Teilstichprobe im Vergleich zur Gesamtstichprobe jedoch nur sehr klein, ist es eher unwahrscheinlich, dass jeder Befragte in der Teilstichprobe vorkommt. Diese Annahmen bezüglich einer Teilabtastung sind sehr unzureichend, denn das alleinige Vorhandensein eines Individuums in einer Stichprobe kann schon negative Folgen mit sich bringen. Somit kann Subsampling ebenfalls nicht als alleiniges Verfahren, sondern nur als ein Baustein, der Differential Privacy betrachtet werden.[117]

Laplace-Mechanismus

Der Laplace-Mechanismus beruht auf dem Hinzufügen von kontrolliertem Rauschen zum Abfrageergebnis, bevor es an den Benutzer zurückgegeben wird. Das Rauschen wird aus der Laplace-Verteilung abgetastet, die bei 0 mit einer Skalierung b zentriert ist. Das Rauschen wird durch Lap(b) dargestellt, wobei ein größeres b ein höheres Rauschen anzeigt. Die entsprechende Wahrscheinlichkeitsdichtefunktion ist:

Laplace-Wahrscheinlichkeitsdichtefunktion:

$$Lap(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right),$$

Laplace-Mechanismus:

$$M(D) = f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

Der Laplace-Mechanismus zeigt, dass die Höhe des Rauschens mit der Sensibilität der Abfrage f und dem Privacy-Budget ϵ zusammenhängt. Das Privacy-Budget ϵ soll das Datenschutz-Level des Laplace-Mechanismus M garantieren. Dabei steht ein schwaches ϵ für ein größeres Datenschutz-Level.[118]

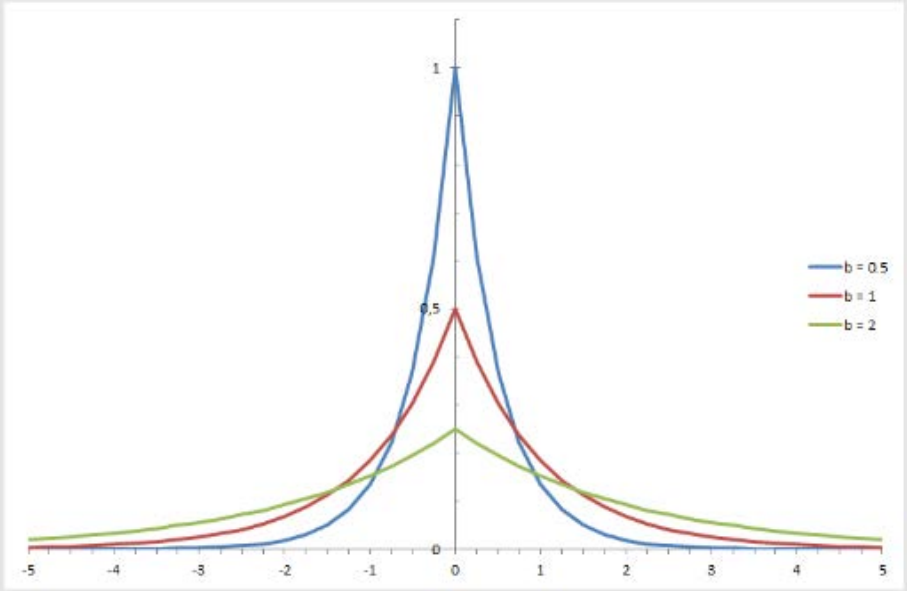

Abbildung 4-8: Lap(x|b) Laplace-Verteilung mit b=0,5; b=1; b=2

117 (Vgl. Dwork 2011, S. 87)
118 (Vgl. Zhu et al. 2017, S. 13)

Wie in der Abbildung 4-8 zu sehen ist, handelt es sich bei der Laplace-Verteilung um eine Doppelexponentialfunktion. Sie berechnet f und durchzieht jede Koordinate der Laplace-Verteilung mit einem Rauschen. Die Skala des Rauschens wird über die Sensibilität f aufgetragen.[119]

Die Sensibilität gibt an, wie sehr Daten eines Individuums das Ergebnis einer Datenbankabfrage beeinflussen können. Denn genau diese personenbezogenen Angaben gilt es bei dem Konzept der Differential Privacy zu schützen. Somit ist die Sensibilität entscheidend dafür, wie viel Rauschen den Daten hinzugefügt werden muss, ohne einen Rückschluss auf das Individuum bekommen zu können. Je höher die Sensibilität ist, desto höher muss das Rauschen sein.

Im Folgenden werden sie Sensibilität und die globale Sensibilität beschrieben.

Die Sensibilität, als zufälliger Fehler Δf ist folgendermaßen definiert:[120]

$$\Delta f = max_{D1,D2}|f(D_1) - f(D_2)|$$

D_1 und D_2 beschreiben die benachbarten Datensätze.

Die Globale Sensibilität wird bei Anfragen mit relativ niedrigen Empfindlichkeitswerten, wie bei Zähl- oder Summenabfragen verwendet. Beispielsweise hat eine Zählerabfrage normalerweise eine globale Sensibilität von f_{GS} von D = 1. Wenn sich die wahren Antworten in einem Bereich von 100 bis 1000 befinden, ist im Gegensatz dazu die Sensibilität sehr niedrig. Für Fragen nach dem Median oder Mittelwert, ergibt die globale Sensibilität sehr hohe Werte im Vergleich zu den wahren Antworten.

Definition:

$$\Delta f_{GS} = max||f(D) - f(D')||$$

Exponential-Mechanismus

Im Gegensatz zu dem Laplace-Mechanismus, wo sich die Anfragen an die Datenbanken auf numerische Werte / Antworten begrenzen, wird der Exponential-Mechanismus für nicht numerische Anfragen eingesetzt. Um dabei die Ergebnisse zu randomisieren, werden sie mit einer Score-Funktion q(D,ϕ) kombiniert. Diese stellt dar, wie gut eine Ausgabe ϕ für einen Datensatz D ist. Die Wahl der Score-Funktion ist abhängig von der Anwendung.

Differential Privacy ist mit Hilfe des Exponential-Mechanismus erfüllt, wenn gilt:

$$M(D) = \{ Rückgabewert\ \phi\ mit\ der\ Wahrscheinlichkeit\ \alpha\ exp(\frac{\epsilon q(D,\phi)}{2\Delta q})\}$$

Mit der Score-Funktion $q(D,\phi)$ von dem Datensatz D, welche die Qualität des Ergebnisses $\phi \in \Phi$ beschreibt. Dabei steht Δq für die Sensibilität der Score-Funktion.

---

119 (Vgl. Dwork und Roth 2014, S. 30 ff.)

120 (Vgl. Zhu 2017, S. 11 ff.)

Beispiel zur Gegenüberstellung von dem Laplace- und Exponential-Mechanismus

In dem folgenden Beispiel soll der Unterschied zwischen dem Laplace-und Exponential-Mechanismus dargestellt werden. Dazu wird der folgende medizinische Datensatz D betrachtet.

Tabelle 4-11: Medizinische Daten mit Personenbezug

Name	Beruf	Geschlecht	Alter	Diagnose
Hans	Bäcker	M	57	Diabetes
Klaus	Ingenieur	M	60	Heuschnupfen
Adam	Anwalt	M	35	Erkältung
Melanie	Lehrerin	W	40	Haarausfall

Tabelle 4-12: Ergebnisse des Exponentialmechanismus

Diagnosen	erkrankte Personen	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 1$
Diabetes	24	0,25	0,32	0,12
Heuschnupfen	8	0,25	0,15	$4x10^{-5}$
Erkältung	28	0,25	0,40	0,88
Haarausfall	5	0,25	0,13	$8,9x10^{-6}$

Hierbei wird die Frage f_1 gestellt: Wie viele Personen in diesem Datensatz leiden an Diabetes? Da die Antwort numerisch ist, kann der Laplace-Mechanismus verwendet werden, um Differential Privacy zu gewährleisten. Als erstes wird dazu im Folgenden die globale Sensibilität analysiert. Dabei wirkt sich das Löschen eines Datensatzes in D maximal um 1 aus, somit ist die Sensibilität von D = 1. Als nächstes wird das Privacy-Budget ϵ für den Laplace-Mechanismus bestimmt, hier mit ϵ = 1. Anschließend wird das Rauschen der wahren Antwort hinzugefügt und der Laplace-Mechanismus mit der verrauschten Antwort M(D)= f_1 (D) + Lap(1) ausgegeben.

Wird nun eine zweite Fragenstellung f_2- „Welche Diagnose kommt am häufigsten vor?“ – gestellt. Wird die Antwort kein numerisches Ergebnis liefern und somit muss der Exponential-Mechanismus verwendet werden. Die oben abgebildete Tabelle 4-12 zeigt alle Diagnosen mit der Anzahl der erkrankten Personen auf. Als erstes wird die Score-Funktion von f_2 definiert. Danach wird die Anzahl der Personen auf jede Diagnose als Score-Funktion q abgebildet. Da das Löschen einer Person eine maximale Auswirkung von 1 auf das Ergebnis von q hat, ist die Sensibilität von q = Δq = 1. Die Wahrscheinlichkeit der Ausgabe wird mit Hilfe des Exponential-Mechanismus berechnet. Tabelle 4-12 zeigt die Ergebnisse für ϵ = 0, ϵ = 0,1 und ϵ = 1. Dabei bedeutet ϵ=0, dass hier der Mechanismus eine Antwort aus den vier Optionen wählt. Diese Ausgabewahrscheinlichkeiten sind bei allen Optionen gleich. Zwar bietet ϵ = 0 das höchste Datenschutzlevel, verliert aber fast alle Funktionen. Wird ϵ = 0,1 angenommen, ist eine „Erkältung“, die Diagnose, die am wahrscheinlichsten getroffen wird und „Haarausfall“, die Diagnose, die am seltensten vorkommt. Dabei ist die Lücke nicht sehr groß, dadurch kann ein akzeptables Maß der Privatsphäre und des Nutzens gewährleistet werden. Wird ϵ = 1 angenommen, ist die Wahrscheinlichkeits-Lücke zwischen „Haarausfall“ und den anderen Diagnosen signifikant. Somit besitzt der Mechanismus einen hohen Nutzen, aber ein niedriges Datenschutzniveau.[121]

[121] (Vgl. Zhu 2017, S. 14 ff.)

Probleme

In der nennenswerten Literatur sind aktuell keine Angriffe und Einschränkungen bei der Verarbeitung dokumentiert.

4.3.4.5 SLICING

Beschreibung

Ein weiterer Lösungsansatz in diesem Zusammenhang kann Slicing[122] sein. Beim Slicing werden - vereinfacht dargestellt - die Daten in Teilrelationen sowohl horizontal als auch vertikal zerlegt, permutiert und schließlich wieder zu einer Relation zusammengesetzt.[123]

Tabelle 4-13: anonymisierte Tabelle vor Slicing

#	Geschlecht	Geburtsjahr	PLZ	Krankheit	#
1	M	1960	44141	Heuschnupfen	1
2	W	1960	35755	Akne	2
3	M	1961	46284	Heuschnupfen	3
4	W	1964	34323	Diabetes	4
5	W	1962	82098	Akne	5
6	M	1964	45117	Heuschnupfen	6
7	W	1960	72465	Diabetes	7
8	M	1964	83098	Diabetes	8
9	W	1963	10223	Heuschnupfen	9
10	W	1962	02387	Akne	10
11	M	1963	23243	Diabetes	11
12	M	1962	75656	Akne	12
13	W	1960	54677	Akne	13
14	M	1961	19485	Diabetes	14
15	M	1964	96354	Akne	15
16	W	1960	89347	Diabetes	16
17	M	1963	58475	Diabetes	17
18	W	1962	74885	Akne	18
19	M	1961	06346	Heuschnupfen	19
20	W	1961	47743	Heuschnupfen	20
21	M	1960	19948	Diabetes	21

122 (Vgl. Li et al. 2009, S. 561 ff.)
123 (Vgl. Grunert und Heuer 2015, S. 24 ff.)

Im Beispiel in der Tabelle 4-13 sind die Daten vertikal in 7er-Pakete und horizontal in die Spaltengruppen Geschlecht und Geburtsjahr sowie PLZ und Krankheit aufgeteilt. Zur Verdeutlichung sind die einzelnen Tupel mit Nummern versehen. Im Anschluss daran können Verschiebungen vorgenommen werden.

In Tabelle 4-14 wurde für die Spaltengruppe Geschlecht und Geburtsjahr in jedem der 7er-Pakete eine Verschiebung um zwei Datensätze nach oben durchgeführt. Diese einfache Verschiebung reicht bereits aus, um eine Zuordnung der Krankheit zu einer Person zu verhindern.

Tabelle 4-14: anonymisierte Tabelle nach Slicing

#	Geschlecht	Geburtsjahr	PLZ	Krankheit	#
3	M	1961	44141	Heuschnupfen	1
4	W	1964	35755	Akne	2
5	W	1962	46284	Heuschnupfen	3
6	M	1964	34323	Diabetes	4
7	W	1960	82098	Akne	5
1	M	1960	45117	Heuschnupfen	6
2	W	1960	72465	Diabetes	7
10	W	1962	83098	Diabetes	8
11	M	1963	10223	Heuschnupfen	9
12	M	1962	02387	Akne	10
13	W	1960	23243	Diabetes	11
14	M	1961	75656	Akne	12
8	M	1964	54677	Akne	13
9	W	1963	19485	Diabetes	14
17	M	1963	96354	Akne	15
18	W	1962	89347	Diabetes	16
19	M	1961	58475	Diabetes	17
20	W	1961	74885	Akne	18
21	M	1960	06346	Heuschnupfen	19
15	M	1964	47743	Heuschnupfen	20
16	W	1960	19948	Diabetes	21

Probleme

Bei diesem Verfahren muss beachtet werden, dass die Informationen an sich verfälscht werden können. So zeigt das Beispiel in Tabelle 4-14, dass eine Anfrage, wie viele im Jahr 1960 geborene Personen an Akne leiden, nicht mehr das korrekte Ergebnis liefert. In der Tabelle 4-14, in welcher Slicing angewendet wurde, wird als Ergebnis der Anfrage nur ein Datensatz ermittelt, während in der ursprünglichen Tabelle 4-13 zwei Datensätze existieren.

5 ANALYSE: VERHINDERUNG VON DE-ANONYMISIERUNG

Bei der De-Anonymisierung werden Daten, die in anonymer oder anonymisierter Form vorliegen, wieder konkreten Personen zugeordnet. Durch die Kombination von verfügbaren Daten wird eine individuelle Bezugnahme ermöglicht, so dass die Zuordnung zu dieser Person sehr wahrscheinlich oder sogar sicher ist.

Aktuell ist dies problematisch, indem veröffentlichte Bilddateien genutzt werden können, um Benutzer sozialer Netzwerke zu de-anonymisieren oder Beziehungen zwischen Nutzern außerhalb des sozialen Netzwerks aufzudecken.[124] Auch durch eine Auslesung des Browserverlaufs und den Abgleich der Daten mit Profilen in sozialen Netzwerken sowie durch den Abgleich verschiedener Datenbanken konnten in der Vergangenheit De-Anonymisierungen durchgeführt werden.[125]

Für den Statistikbereich hat das BVerfG im "Volkszählungsurteil" festgelegt, dass zur Sicherung des Rechts auf informationelle Selbstbestimmung Löschungsregelungen für persönliche Identifikationsmerkmale erforderlich sind, die ansonsten eine De-Anonymisierung leicht ermöglichen würden. Zum Schutz des Rechts auf informationelle Selbstbestimmung sei eine möglichst frühzeitige faktische Anonymisierung durchzuführen, verbunden mit Vorkehrungen gegen eine De-Anonymisierung.[126]

Auch die DSGVO greift dies auf, indem sie in Erwägungsgrund 47 und 50 darauf abstellt, dass die betroffenen Personen u.a. aufgrund einer vorgenommenen Anonymisierung nicht mit einer Identifikation rechnen müssen. Die De-Anonymisierung erscheint damit als besonders schwere Beeinträchtigung ihrer Rechte.

Eine De-Anonymisierung von Daten kann nur dann sicher verhindert werden, wenn strikte Regeln in organisatorischer, technischer und juristischer Hinsicht eingehalten werden.

5.1 ANFORDERUNGEN AN ANONYMISIERUNG

Im Folgenden sollen die in Kapitel 4.3.4 vorgestellten Verfahren zur Anonymisierung mit einer Nutzwertanalyse auf ihre Tauglichkeit überprüft werden. Im Rahmen der Nutzwertanalyse werden die Interessen der betroffenen Personen sowie der auswertenden Stelle berücksichtigt. Somit findet eine Differenzierung der Gewichtung in den Bereichen *Schutz des Einzelnen*, begründet durch das Recht auf informationelle Selbstbestimmung (Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG), sowie *Datenqualität* statt. Da dem Schutz des Einzelnen eine höhere Bedeutung zugewiesen wird, wird der Schutz des Einzelnen mit 60% und die Datenqualität mit 40% gewichtet.

Die Gewichtung im Rahmen des Schutzes des Einzelnen wird weiterhin in Identity Disclosure, Attribute Disclosure und Membership Disclosure aufgeteilt.

---

124 (Vgl. Greveler und Löhr 2010, S. 1 f.)
125 (Vgl. Schmitz 2010, S. 6 ff.)
126 (BVerfG, Urt. v. 15.12.1983, 1 BvR 209/83, unter C.IV.4.b)

Zu den Anforderungen, die zum Schutz der einzelnen Personen dienen, wurde die Vermeidung von Identity Disclosure mit einer Wichtigkeit von 30% eingestuft, dies resultiert daraus, dass eine mögliche Re-Identifikation einer Person den höchsten Schaden für diese herbeiführen kann.

Eine 20%ige Gewichtung gilt der Vermeidung von Attribute Disclosure, wobei hier im Gegensatz zur Identity Disclosure, nur die Möglichkeit der Zuordnung einer Person zu einem bestimmten Attribut bestehen kann.

Weiterhin wird der Membership Disclosure eine Gewichtung von 10% zugeordnet, dies ist dadurch begründet, dass es hierbei im Gegensatz zu der Identity-und Attribute-Disclosure nur möglich ist, zu identifizieren, ob sich diese Person allgemein in diesem Datensatz befindet.

Die folgende Tabelle zeigt auf, wie sich die unterschiedlichen Anonymisierungsverfahren zu den definierten Anforderungen verhalten:

Tabelle 5-1: Zusammenfassung der Nutzwertanalyse

Anforderungen / Verfahren	Verhinderung Identity Disclosure	Verhinderung Attribute Disclosure	Verhinderung Membership Disclosure	Datenqualität
Gewichtung	30%	20%	10%	40%
k-Anonymität	3	1	1	3
l-Diversity	3	2	1	3
t-Closeness	3	3	3	3
Differential Privacy	3	3	3	2
Slicing	3	3	3	1

Die Schutzreserve wurde in der Nutzwertanalyse nicht berücksichtigt, da sie bei allen genannten Methoden gleich hoch ist. Eine explizite Berücksichtigung der Schutzreserve bei der Bewertung der einzelnen Anonymisierungsverfahren findet nicht statt, da diese abhängig von der Konfiguration der jeweiligen Verfahren ist.

Das Anonymisierungsverfahren der k-Anonymität schützt vor Identity Disclosure, hier mit der Gewichtung 3 angenommen, jedoch besteht kein Schutz vor Attribute-und Membership Disclosure. Deswegen wurden diese Verfahren wie in der Tabelle zu erkennen ist nur mit dem Faktor 1 gewichtet. Wird die Datenqualität der k-Anonymität betrachtet, zeigt diese keinen Verlust auf. Somit wird ihr der Faktor 3 zugeordnet.

l-Diversity verhindert ebenfalls die Re-Identifikation von Personen zu bestimmten Datensätzen, sodass der Identity Disclosure die Gewichtung 3 zugeordnet werden kann. Zwar bietet die l-Diversity einen höheren Schutz von Attribute Disclosure als die k-Anonymität, jedoch kann, wie in Kapitel 4.3.4.2 beschrieben, durch die Skewness und Similarity-Attack kein ausreichender Schutz gewährleistet werden. Somit wurde die Verhinderung von Attribute Disclosure in der Tabelle mit der Gewichtung von 2 angenommen. Die Vermeidung der Membership-Disclosure wird bei der l-Diversity nicht erreicht und wird deswegen mit dem Faktor 1 bewertet. In Bezug auf die Qualität der Daten zeigt die l-Diversity keine Nachteile auf. Aus diesem Grund wurde sie in der Tabelle mit dem Faktor 3 gewichtet.

t-Closeness bietet einen Schutz vor Identity und Attribute Disclosure und Membership Disclosure. Deswegen wurden diese Anforderungen in der Tabelle auch mit dem Faktor 3 gewichtet. Zusätzlich zeigt die Qualität der Daten auch keine Nachteile auf.

Differential Privacy kann eine Reidentifikation von Daten verhindern und schützt zudem vor Attribute- und Membership Disclosure. Somit werden diese Anforderungen auch mit dem Faktor 3 bewertet. Weiterhin entspricht Differential Privacy den genannten Anforderungen in Bezug auf die Datenqualität, hierbei muss jedoch gewährleistet sein, dass durch das hinzugefügte Laplace-Rauschen im Datenbestand nur ein geringer Informationsverlust stattfinden darf. Deswegen wurde die Anforderung an die Datenqualität hierbei nur mit dem Faktor 2 bewertet.

Slicing als Anonymisierungsverfahren kann die Anforderungen zum Schutz der Individuen, wie Identity-, Attribute- und Membership Disclosure erfüllen. Somit wurden diese Anforderungen in der Tabelle mit dem Faktor 3 gewichtet. Jedoch werden durch das Slicing die Datenbestände so verändert, dass die Daten keine brauchbaren Informationen mehr liefern können. Die Gewichtung wird hierbei mit dem Faktor 1 angenommen.

In Bezug auf die Nutzwerte der einzelnen Anonymisierungsverfahren ergeben sich aus der oben abgebildeten Tabelle folgende Ergebnisse: Das Slicing hat mit 2,2 den niedrigsten Nutzwert für die Anonymisierung ergeben. Weiterhin wurde die k-Anonymität mit einem Nutzwert von 2,4 berechnet und die l-Diversity mit einem Nutzwert von 2,6. t-closeness bietet mit 3,0 den höchsten Nutzwert und ist daher für die Anonymisierung der Daten zu priorisieren. Die Priorisierung des Verfahrens t-closeness ist vor allem gegenüber dem Verfahren Differential Privacy (Nutzwert von 2,6) dadurch zu erklären, dass die Datenqualität, die für die angestrebte Auswertung der Daten sichergestellt sein muss, durch Differential Privacy reduziert werden kann. Zwar bietet Differential Privacy den höchsten Schutz vor De-Anonymisierung, jedoch muss – ebenso wie beim Slicing – die Beeinträchtigung der Datenqualität in die Betrachtung gezogen werden.

Die Bewertung der einzelnen Anonymisierungsverfahren auf Grundlage der oben dargestellten Nutzwertanalyse zeigt, dass die technischen Verfahren zur Anonymisierung personenbezogener Daten vorhanden sind. Ebenso wird durch die Erläuterungen der jeweiligen Verfahren (siehe dazu Kapitel 4.3.4) deutlich, dass die Verfahren nur dann eine akzeptable Anonymisierung gewährleisten können, wenn die notwendigen Parameter entsprechend einem hohem Anonymisierungsgrad ausgewählt werden. So sind neben den technischen Verfahren weitere organisatorischen Maßnahmen zu ergreifen, um eine hinreichende Anonymisierung sicherzustellen.

5.2 EINSATZ EINES DATENTREUHÄNDERS

Durch den Einsatz eines Datentreuhänders kann, wenn die im weiteren Verlauf des Gutachtens aufgeführten Regeln (siehe Kapitel 5.5) eingehalten werden, eine De-Anonymisierung verhindert werden. Über eine Public-Key-Infrastruktur könnte ein Datentreuhänder (Trust Center) Signaturen vergeben. Hierdurch kann einerseits eine eindeutige Identifizierbarkeit der Quelle einer übertragenen Information gewährleistet werden. Anderseits wird auf diese Weise ausgeschlossen, dass personenbezogene Informationen, die nur anhand der Signatur persönlich zugeordnet werden können, veröffentlicht werden.

Als Beispiel für eine Datentreuhänderschaft kann die Kooperation von Microsoft und der Telekom-Tochter T-Systems dienen. Microsoft bietet Dienste wie Office 365, Dynamics online und Azure in

der Cloud an. T-Systems tritt als Datentreuhänder auf und kontrolliert den Zugriff auf sensible Daten. Somit wird garantiert, dass Daten nicht unerlaubt an Dritte weitergegeben werden.

Da in o.g. Beispiel der Hauptzweck aber in der operativen Verarbeitung von Daten liegt, ist hier der Anonymisierungsaspekt nicht gegeben. Zudem ist die o.g. Kooperation auch nicht unmittelbar auf ein allgemeines Treuhändermodell übertragbar, da T-Systems selbst als wirtschaftliches Unternehmen finanzielle Interessen verfolgt, ein hinreichend zuverlässiger Datentreuhänder aber unabhängig handeln muss. Um eine De-Anonymisierung sicher verhindern und gleichzeitig einen Zugriff der Berechtigten garantieren zu können, müssen weitergehende Anforderungen in organisatorischer, rechtlicher und technischer Hinsicht an einen Datentreuhänder gestellt werden.

5.3 DER EINSATZ VON DATENTREUHÄNDERN IN DER PRAXIS

Es gibt einige Beispiele aus der Praxis, in denen Datentreuhänder eingesetzt werden.

Die Idee, einen Datentreuhänder einzusetzen, wurde schon im Jahr 2007 aufgeworfen: „Im Hinblick auf die Realität umfassender Sammlung, Speicherung und Verarbeitung von Proben und Daten in Biobanken reicht es möglicherweise nicht aus, allein mithilfe des Instruments der informierten Einwilligung einen genügenden Persönlichkeitsschutz zu generieren. [...] Modellcharakter könnte hier die Institution eines Treuhänders für Biobanken gewinnen.“[127] Mittlerweile sind einige dieser Biobanken eingerichtet, exemplarisch seien hier die Westdeutsche Biobank Essen (WBE) und die Zentrale Biobank Tübingen genannt.

In §12a Abs. 4 des Hamburgischen Krankenhausgesetzes wird die Einsetzung eines Datentreuhänders vorgeschlagen: „Bei einer Verwendung der Sammlung zu genetischer Forschung ist zu prüfen, ob die Sicherheit der betroffenen Personen vor einer unbefugten Zuordnung ihrer Proben und Daten es erfordert, dass die Pseudonymisierung nach den Absätzen 2 und 3 durch eine unabhängige externe Datentreuhänderin oder einen unabhängigen externen Datentreuhänder erfolgt.“[128]

Im Verkehrsbereich können als Beispiele die Verkehrsinfrastrukturfinanzierungsgesellschaft (VIFG), die FSD GmbH - Zentrale Stelle nach StVG - oder die österreichische Autobahnen- und Schnellstraßen-Finanzierungs-Aktiengesellschaft (ASFINAG) dienen.

Nicht zuletzt können auch die Krebsregister in Deutschland als eine Art Datentreuhänder angesehen werden. So teilen sich die Krebsregister in zwei Bereiche auf – zum einen den Vertrauensbereich und zum anderen den Registerbereich. Der Vertrauensbereich beinhaltet die Datenannahmestelle, in welcher die Daten nach der Übermittlung durch die meldepflichtigen Personen pseudonymisiert werden. Die Speicherung der personenidentifizierbaren Daten inkl. Zuordnungsregel erfolgt im Vertrauensbereich. Der Registerbereich erhält nach der Pseudonymisierung nur noch die Registerdaten (medizinischen Daten über die Krebsbehandlung) sowie das Pseudonym. Anfragen von Dritten hinsichtlich der Datenherausgabe für Forschungsfragen werden durch den Vertrauensbereich bearbeitet und bei erfolgreicher Prüfung in anonymisierter Form an die anfragende Stelle übermittelt.[129]

---

127 (Revermann und Sauter 2007, S. 179)
128 (HmbKHG, § 12a Abs. 4)
129 (Vgl. Gesetz über das Klinische Krebsregister Niedersachsen (GKKN). Insbesondere die §§ 10, 11 und 20)

Diese Einrichtungen haben bereits gezeigt, dass ein Kompetenzzentrum für Infrastrukturinvestitionen enorme Vorteile bieten kann. Es ist unter anderem in der Lage, einheitliche Strukturen zu entwickeln und Standards zu setzen, Bewertungen zu professionalisieren und zu vereinheitlichen und einen einheitlichen, kompetenten Ansprechpartner bereitzustellen.

Die österreichische Autobahnbetreibergesellschaft ASFINAG beispielsweise wurde mit Kapital vom Bund ausgestattet. Sie sammelt und stellt bereits heute sämtliche Verkehrsdaten über ihre zentrale Datendrehscheibe bereit. Darüber hinaus beteiligt sie sich an verschiedenen Projekten zur Planung und Erforschung der Verkehrsinfrastruktur und hat hierzu u.a. bereits auf dem Testfeld Telematik eine Vielzahl von Sensoren eingerichtet, um eine effiziente Erfassung von Verkehrsdaten zu ermöglichen.

5.4 PSEUDONYMISIERUNG UND ANONYMISIERUNG BEIM DATENTREUHÄNDER

Der Datentreuhänder setzt sowohl Verfahren zur Pseudonymisierung als auch zur Anonymisierung von personenbezogenen Daten ein. Eine Pseudonymisierung personenbezogener Daten erfolgt nach Übermittlung der Daten durch die ursprüngliche, datenverarbeitende Stelle. Durch eine Pseudonymisierung der Daten besteht für den Datentreuhänder jederzeit die Option, weitere Attribute einer Person bzw. Pseudonym zu zuordnen. Die Anonymisierung von Daten findet vor der Ausgabe der Daten an auswertende Dritte statt - somit ist es diesen gar nicht bzw. nur mit erheblichem Aufwand an monetären und zeitliche Ressourcen möglich, die Daten zu deanonymisieren.

Die Abbildung 5-1 zeigt den Fluss der personenbezogenen Daten von einer erhebenden Stelle zum Datentreuhänder, die Verarbeitung innerhalb des Treuhänders sowie die Ausgabe der anonymisierten Daten an eine auswertende Stelle. Das Teilkapitel 5.4.2 beschreibt die Vorgehensweise der Datenübermittlung sowie -verarbeitungen detailliert.

Vor der Speicherung der Daten beim Datentreuhänder muss eine Trennung der Daten in Vertrauensdaten (Pseudonyme und persönliche Daten) und Nutzdaten erfolgen, die mit Pseudonymen abgespeichert werden. Dies ist notwendig, um eine Fortschreibung der Daten im Zeitablauf gewährleisten zu können. Als praktisches Beispiel kann die Entwicklung eines Krankheitsverlaufs angeführt werden: Eine Speicherung von anonymisierten Daten bei der Diagnose der Krankheit würde dazu führen, dass der weitere Krankheitsverlauf nicht der jeweiligen Person zugeordnet werden kann und somit Forschungsfragen bzgl. der Wirksamkeit von Medikamenten oder Behandlungen nicht beantwortet werden können.

Auch eine Generalisierung von beispielsweise dem Geburtsdatum in eine Altersangabe oder eine ungefähre Altersangabe ist nicht zielführend, da die Daten bzgl. des Alters in Zukunft nicht mehr ausgewertet werden können. Vor der Weitergabe der Daten an Dritte wird eine spezielle Anonymisierung jeweils mit Bezug auf die unterschiedlichen Anforderungen, die sich aus den Forschungsfragen ergeben, durchgeführt.

Zusätzlich muss dafür Sorge getragen werden, dass die Daten so abgespeichert sind, dass ein Zugriff von unberechtigten Dritten nicht möglich ist. Dabei ist dafür Sorge zu tragen, dass die Speicherung der Daten sowohl in Bezug auf die Datenbanksysteme, als auch in Bezug auf die Verschlüsselung stets auf dem aktuellen Stand der Technik zu halten ist. Es sollten regelmäßig Penetrations-Tests durch unabhängige Spezialisten durchgeführt werden, um Angriffen von außen weitestgehend zu verhindern bzw. wesentlich zu erschweren.

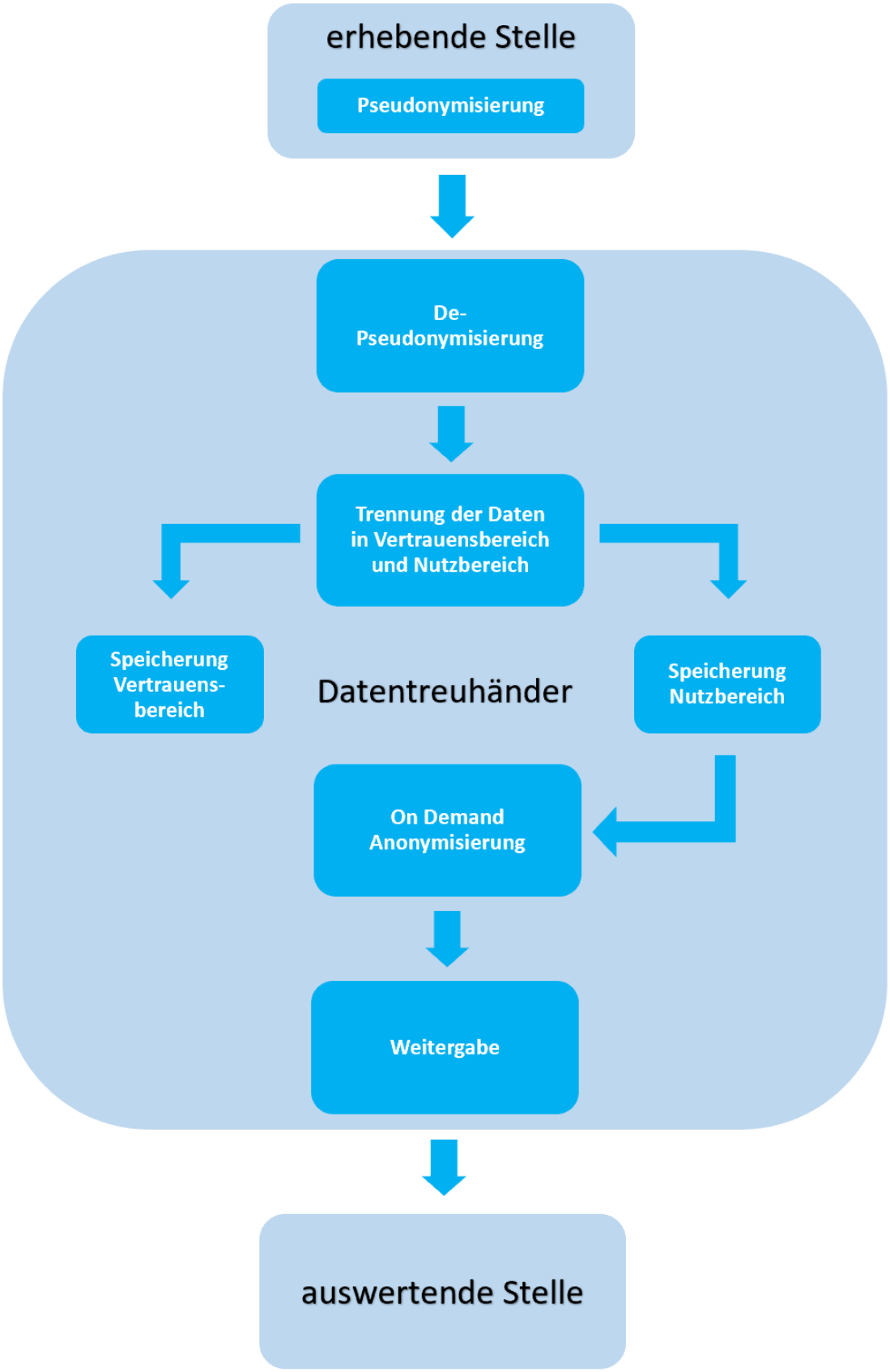

Abbildung 5-1: Prozess beim Datentreuhänder

5.5 VORGEHENSWEISE

Im Nachfolgenden werden die Verarbeitungstätigkeiten zwischen den Akteuren - erhebende Stelle, Datentreuhänder und auswertende Stelle - beschreiben.

Übertragung der Daten an den Datentreuhänder

Die Daten werden durch den Datentreuhänder von der erhebenden Stelle angefordert oder evtl. automatisiert von der erhebenden Stelle an den Datentreuhänder übermittelt. Die erhebende Stelle pseudonymisiert die Daten vor der Übermittlung. Die Übermittlung der Nutzdaten wird getrennt von der Übermittlung der Vertrauensdaten durchgeführt, damit im Fall einer Penetration der Datenübertragung eine De-Pseudonymisierung nicht möglich ist. Die Übertragung erfolgt jeweils in verschlüsselter Form mit einer dem aktuellen Stand der Technik entsprechenden hohen Verschlüsselungsstärke.

De-Pseudonymisierung und Speicherung beim Datentreuhänder

Der Datentreuhänder de-pseudonymisiert im ersten Schritt die Daten, sodass die ursprünglichen Informationen in reiner Form vorliegen. In einem zweiten Schritt führt der Datentreuhänder eine eigene Pseudonymisierung durch oder ordnet die übermittelten Daten einem existierenden Pseudonym zu, wenn es sich um fortgeschriebene Daten handelt. Liegt keine Fortschreibung von Daten vor, erfolgt im dritten Schritt eine Trennung in Vertrauensdaten (Pseudonym und persönliche Daten) und Nutzdaten (Pseudonyme und Nutzdaten). Diese Daten werden jeweils getrennt voneinander in eigenen Datenbanken gespeichert. Die Speicherung erfolgt ebenfalls mit einer dem aktuellen Stand der Technik hohen Verschlüsselung, um im Falle eines Datendiebstahls die Anonymität der Daten gewährleisten zu können. Eine ähnliche Vorgehensweise (Trennung in Vertrauensdaten und Nutzdaten) wird bereits im Krebsregister angewendet.

Anfrage an den Datentreuhänder und Datenweitergabe

Wird eine Anfrage für eine bestimmte Forschungsfrage an den Datentreuhänder gestellt, stellt dieser die entsprechenden Daten so zusammen, dass eine De-Anonymisierung nicht möglich ist, d.h. alle nicht notwendigen Merkmale werden entfernt oder soweit wie möglich vergröbert. Ist beispielsweise das Geschlecht für eine Auswertung nicht von Belang, so wird diese Information für die Übertragung zum Dritten entfernt. Ist das genaue Alter nicht wichtig, so werden Altersklassen gebildet.

Dazu muss beachtet werden, dass lediglich die Werte übermittelt werden, die für die Forschungsfrage von Relevanz sind. Weiterhin muss dafür Sorge getragen werden, dass die in Kapitel 5.5 genannten Regeln eingehalten werden.

Sind die o.g. Voraussetzungen erfüllt, überträgt der Datentreuhänder die Informationen an den Auftraggeber in verschlüsselter Form. Hierbei werden die Informationen über den Auftraggeber, den Zweck der Anfrage, das Anfragedatum, die Pseudonyme der übertragenen Datensätze und die Metadaten protokolliert, um im Nachhinein nachweisen zu können, dass die Übertragung der Informationen dem Zweck angemessen war.

Der Datentreuhänder hat sicherzustellen, dass durch wiederholte Anfragen eine De-Anonymisierung durch Überschneidungs- bzw. Hintergrundwissen nicht möglich ist. Dazu ist für jede Anfrage eine Protokollierung der übertragenen Metadaten durchzuführen. Durch die Protokollierung der Anfragen

und der damit in Verbindung stehenden übertragenen Informationen ist eine rechtzeitige Erkennung von riskanten Übertragungen möglich.

5.6 ANONYMISIERUNGSVERFAHREN BEIM DATENTREUHÄNDER

In Kapitel 4.3.4 wurden unterschiedliche Verfahren zur Anonymisierung von personenbezogenen Daten sowie die jeweiligen verfahrensspezifischen Probleme erörtert. Die benannten Probleme der einzelnen Anonymisierungsverfahren müssen bei ihrer Anwendung berücksichtigt werden.

In Bezug auf dem für den Datentreuhänder beschriebenen Verfahren (s. Kapitel 5.4.2) müssen die personenbezogenen Daten vor der Ausgabe an einen auswertenden Dritten derart anonymisiert werden, dass auch mit einem sehr umfangreichem Hintergrundwissen oder durch die Verkettung mit anderen Daten eine De-Anonymisierung nicht möglich ist.

Auf Grundlage der Ergebnisse der Nutzwertanalyse in Kapitel 5.1 sind hinsichtlich des Schutzes des Einzelnen die Anonymisierungsverfahren *t-Closeness* (s. Kapitel 4.3.4.3) oder *Differential Privacy* (s. Kapitel 4.3.4.4) zu verwenden. Beide Verfahren bieten ausreichend Schutz gegen Identity Disclosure, Attribute Disclosure und Membership Disclosure.

Wird zusätzlich noch der Aspekt der Datenqualität einbezogen, so ist bei dem Verfahren *Differential Privacy* darauf zu achten, dass durch das Rauschen die Daten nicht derart verfälscht werden, dass die Ergebnisse im Hinblick auf die Forschungsfrage unbrauchbar werden. Das Verfahren *t-Closeness* hat in Bezug auch die Datenqualität keine Einschränkungen.

5.7 ORGANISATORISCHER ASPEKT

Um einem Missbrauch der Funktion als Treuhänder vorzubeugen, muss der Datentreuhänder eine unabhängige Stelle sein, die keine Gewinnerzielungsabsicht verfolgt.

Bei einer oberflächlichen Betrachtung könnte man zunächst annehmen, dass der Treuhänder unter regulierten Voraussetzungen gewinnorientiert am Markt agieren könnte. Diese Ansicht ist jedoch zu verneinen. Dem Treuhänder ist es möglich, auf eine große Menge von personenbezogenen Daten zuzugreifen, bevor eine Anonymisierung erfolgt. Dies bedeutet, dass der Treuhänder Rohdaten besitzt, welche ihn gegenüber anderen Marktteilnehmern privilegiert. Ferner muss angeführt werden, dass der Treuhänder ein Profiteur seiner Stellung als Treuhänder ist. Nur aufgrund seiner besonderen vertrauensvollen Stellung sowie einer absoluten Transparenz werden Unternehmen sowie die öffentliche Verwaltung gewillt sein, ihm die eigenen personenbezogenen Daten zu übermitteln, da sie sich einen eigenen Vorteil (in diesem Fall der Zugriff auf alle relevanten zusammengefügten Daten) erhoffen.

Wenn der Treuhänder nun das Bestreben verfolgen würde, gewinnorientiert zu arbeiten, müsste er permanent nach Möglichkeiten suchen, Gewinne zu erzielen. Die evidenten Möglichkeiten hierfür wären unter anderem eine Kostensenkung, die Erhöhung von Preisen sowie die Erschließung neuer Geschäftsfelder.

Da die Daten auf den höchsten technischen Standards gesichert und anonymisiert werden müssen, liegt in diesem Bereich ein wesentlicher Kostenfaktor. Aufgrund der rasanten technischen

Entwicklung wird in diesem Bereich langfristig keine Kosteneinsparung möglich sein. Die Kosten für eine entsprechende Infrastruktur sowie qualifizierte Mitarbeiter werden tendenziell eher steigen. Hier ist insbesondere anzuführen, dass aufgrund der „Werthaltigkeit" der Rohdaten, die allerhöchsten Sicherheitsstandards eingehalten werden müssen.

Auch die beiden anderen Varianten der Kostensenkungen, die Erschließung neuer Geschäftsfelder sowie die Erhöhung der Preise sind keine ernstzunehmenden Alternativen. Das Modell funktioniert lediglich mit dem Vertrauen in den Treuhänder sowie der Möglichkeit der eigenen Partizipation. Das Produkt (entsprechend generierte anonymisierte Daten) ist abhängig von den Daten, die Unternehmen oder die öffentliche Hand im Vorfeld zur Verfügung stellen. Ab dem Zeitpunkt, in dem die Inhaber der Rohdaten nicht mehr gewillt sind, diese dem Treuhänder zur Verfügung zu stellen, sind die Datensätze des Treuhänders auch nicht mehr aktuell, was sich maßgeblich in der Werthaltigkeit der Daten wiederspiegelt.

Sobald Unternehmen nun einen werthaltigen Bestandteil des Endproduktes liefern (Rohdaten) und Beträge zahlen müssen, welche der Gewinnoptimierung des Treuhänders dienen, wird die Bereitschaft, sich hieran zu beteiligen, sinken. Ferner sollte der Zugriff auf die anonymen Daten grundsätzlich jedem zu einem kostendeckenden Betrag möglich sein. Dies führt dazu, dass neue Analyseansätze vorgenommen werden können und aus den ausgewerteten Daten neue Ideen und Unternehmen entstehen können. Sobald der Zugriff auf die Daten lediglich durch sehr finanzstarke Akteure vorgenommen werden kann, würde durch den Treuhänder eine bestehende Marktmacht manifestiert und Innovationen von kleinen Unternehmen oder Startups verhindert. Eine solche Einschränkung des Wettbewerbs kann nicht im Sinne unserer Wirtschaftsordnung sein.

Auch die Erschließung neuer Geschäftsfelder sollte nicht vorgenommen werden. Das Kerngeschäft des Treuhänders ist die Anonymisierung, Auswertung und individuelle Bereitstellung von Daten. Sobald der Treuhänder nun neue Geschäftsfelder erschließen müsste, würde dies höchstwahrscheinlich in seiner Kernkompetenz, der Datenanalyse und Datenauswertung, geschehen. Hier herrscht jedoch ein enormer Wettbewerbsvorteil gegenüber den anderen Marktteilnehmern. Als Inhaber der Rohdaten können andere Erkenntnisse erlangt und monetarisiert werden, als von Personen oder Unternehmen, welche einen solchen Zugriff nicht haben. Dies führt dazu, dass die „Lieferanten" des Treuhänders, die Unternehmen welche die Rohdaten liefern, sich überlegen werden, ob sie die Daten weiterhin zur Verfügung stellen, um einem „Konkurrenten" am Markt einen Vorteil zu verschaffen.

Aus diesem Grund ist eine Gewinnorientierung des Treuhänders abzulehnen. Er soll mit seinen Daten helfen neue Erkenntnisse, z.B. in dem Gesundheitswesen, der Wissenschaft oder der Wirtschaft, zu erlangen. Aus den oben ausgeführten Gründen wäre eine Gewinnorientierung daher eher kontraproduktiv. Der Treuhänder sollte vielmehr in einer wirtschaftlich gesicherten unabhängigen Institution angesiedelt werden. Ein Beispiel für eine Datentreuhänderschaft ohne Gewinnerzielungsabsicht ist die WBE – Westdeutsche Biobank Essen.

Ein (ggf. beliehener) unabhängiger Datentreuhänder bietet den Vorteil, dass alle relevanten Daten zentral verwaltet, verarbeitet und bereitgestellt werden. Ein hierfür vorzusehendes „Trust Center" nimmt folglich eine Vermittlungsposition zwischen den Datenverantwortlichen, den betroffenen Personen und den berechtigten Dritten ein.

Als vertrauenswürdige Stelle kann sie entsprechende Auskunftsanträge Dritter prüfen und bei berechtigten Auskunftsbegehren die vorliegenden Daten unter Beachtung der Anonymisierung an den

Berechtigten übermitteln. Dadurch wird auch die Gefahr einer möglichen Manipulation der Schnittstellen und der erhobenen Daten vermieden.

Denkbar ist, den Datentreuhänder als staatlich beliehene Stelle zu definieren. Grundlage der Beleihung wäre die Erfüllung hoheitlicher Aufgaben, die aufgrund kollidierender Eigeninteressen nicht von Privatunternehmen in eigener Verantwortung ausgeführt werden können, da deren Eigeninteressen nicht immer mit den staatlich geschützten Interessen der Allgemeinheit übereinstimmen. Der Staat setzt hier für die Vorgaben einen geeigneten Rahmen. Das beliehene Unternehmen wird nicht im (wirtschaftlichen) Eigeninteresse tätig, sondern im Rahmen einer staatlichen Steuerung zur Einhaltung der Vorgaben. Indem das Unternehmen die entscheidenden (rechtlichen) Vorgaben in technische Vorgaben umsetzt, übt es eine hoheitliche Tätigkeit aus.

Der Realisierung von Finanzierungsgesellschaften sind jedoch durch das Demokratieprinzip nach Art. 20 Abs. 2 GG rechtliche Grenzen gesetzt, da jede Ausübung von Staatsgewalt zumindest mittelbar demokratisch legitimiert sein muss. Dies bedeutet, dass bei der Einrichtung ein Verbleib der Aufgabe in öffentlicher Hand oder eine vertragliche Sicherung entsprechender Kontroll- und Einwirkungsmöglichkeiten beachtet werden muss.

In Deutschland könnte daher eine Finanzierungsgesellschaft gegründet werden, der im Interesse des Bundes konkrete Aufgaben zur effektiven Planung und Einrichtung intelligenter Datentreuhandeinrichtungen übertragen werden.

Entsprechend dem österreichischen Modell könnte darüber hinaus frühzeitig eine Schlichtungsstelle zur außergerichtlichen Streitbeilegung eingerichtet werden. Durch die ständige Entwicklung neuer Dienste und Anwendungen im Bereich Big Data muss die Einhaltung entsprechender Datenschutz- und Datensicherheitsstandards gewährleistet sein. Mithin ist ein reibungsloser Geschäftsablauf zwischen den Bereitstellern von entsprechenden Diensten und deren Kunden im B2B-Bereich (Business to Business) und im B2C-Bereich (Business to Consumer) von immenser Bedeutung. Ziel des Schlichtungsverfahrens sollte es sein, bei Streitigkeiten in diesen Bereichen innerhalb eines angemessenen Zeitraumes ein für alle Beteiligen akzeptables Ergebnis herbeizuführen, um so kostspielige und langwierige Prozesse zu vermeiden.

Eine Finanzierung des Datentreuhänders kann u.a. durch den Infrastrukturfonds, in den die durch Nichtbeachtung der EU DS-GVO anfallenden Bußgelder einfließen, erfolgen. Zudem sollten Finanzierungsbeiträge der nutzenden Unternehmen, aber auch staatliche Zuschüsse vorgesehen werden, um die Interessen aller Beteiligten an der treuhänderischen Dateninfrastruktur in allen Bereichen (z.B. Verkehr, Gesundheitswesen, Marktforschung) abzubilden.

Auch eine Stellung des Datentreuhänders als Stiftung ist denkbar, wegen der relativ strengen rechtlichen Vorgaben aber weniger realistisch.

Der Datentreuhänder darf inhaltlich keinesfalls weisungsgebunden sein, weder von staatlicher, noch von Unternehmensseite aus. Die Installation eines unabhängigen Kontrollgremiums ist unerlässlich.

Die Prozesse, die in Zusammenhang mit der Speicherung von Daten in anonymisierter Form und mit der Weitergabe der Daten an Dritte stehen, müssen transparent sein. Sämtliche Transaktionen, die zu einer Speicherung von Daten oder zu einer Weitergabe von Daten führen, müssen nachvollziehbar protokolliert werden. Eine Protokollierung einer jeden Anfrage ist obligatorisch, wobei das Datum der Anfrage, der Zweck der Datenauswertung und der Auftraggeber zwingend gespeichert werden

müssen. Ebenso ist zu speichern, welche Daten übermittelt wurden. Hierbei sind das Identifikation in pseudonymisierter/anonymisierter Form und die Metadaten der Übermittlung zu protokollieren.

Regelmäßige Überprüfungen durch unabhängige Stellen wie z.B. Datenschutzbehörden müssen erfolgen.

Nach Art. 5 Abs. 1 lit. f EU DS-GVO gilt der Grundsatz der Integrität und Vertraulichkeit. Entsprechend dieser Vorschrift müssen personenbezogene Daten „in einer Weise verarbeitet werden, die eine angemessene Sicherheit der personenbezogenen Daten gewährleistet, einschließlich Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung durch geeignete technische und organisatorische Maßnahmen […]". Integrität bedeutet Unversehrtheit, Unverfälschtheit und Vollständigkeit der Daten, während die Vertraulichkeit bedeutet, dass die die Daten quantitativ und qualitativ vor einem fremden Zugriff gesichert werden müssen.[130] Hieraus und aus Art. 5 Abs. 2 EU DS-GVO (Rechenschaftspflicht, d.h. der Verantwortliche muss die Einhaltung der datenschutzrechtlichen Grundsätze darlegen) kann abgeleitet werden, dass die Mitarbeiter des Datentreuhänders auf die Vertraulichkeit verpflichtet werden müssen.

Die für die Ermittlung der für eine konkrete Forschungsfrage zuständigen Mitarbeiter des Datentreuhänders müssen sowohl in technischer Hinsicht, als auch in organisatorischer Hinsicht sehr gut qualifiziert sein und regelmäßig geschult werden. Möglicherweise ist – um Missbrauch verhindern zu können – ein Vier-Augen-Prinzip bei der Abfrage der Daten sinnvoll. Dies bedeutet, dass zwei Personen durch Eingabe eines Kennworts oder eine andere Authentifizierungsmethode die Abfrage freischalten. Hier muss allerdings geprüft werden, ob der Aufwand gerechtfertigt ist, da davon ausgegangen werden kann, dass die Mitarbeiter des Treuhänders vertrauenswürdig sind. Eine Überprüfung der Qualifikation der Mitarbeiter durch staatliche Stellen muss jederzeit - auch ohne Ankündigung - zugelassen werden und möglich sein.

Die Zugriffe auf die Daten müssen protokolliert werden. Die Protokolle müssen vor unberechtigter Änderung bzw. Löschung geschützt werden.

Auch im administrativen Bereich müssen qualifizierte Mitarbeiter eingesetzt werden, die sowohl das technische Knowhow besitzen, um die Sicherheit der Daten vor Diebstahl, Verlust und Zerstörung sicherstellen können, als auch die Administration und Wartung der Server und Datenbanken beherrschen, so dass eine Beauftragung externer Dienstleister nicht oder nur in Ausnahmefällen erforderlich ist. Ist es notwendig, dass ein externes Unternehmen Wartungs- oder Reparaturarbeiten bzw. einen Austausch von Hardware durchführen muss, so sollte sichergestellt werden, dass die Mitarbeiter des externen Dienstleisters zu keiner Zeit allen in den Räumlichkeiten arbeiten, damit ein missbräuchlicher Zugriff auf die Daten durch den Dienstleister nicht erfolgen kann.

Es sind Maßnahmen zu definieren und transparent zu kommunizieren, falls ein Mitarbeiter gegen organisatorische Vorgaben verstößt.

Der Datentreuhänder muss sicherstellen, dass eine De-Anonymisierung auf der Basis der übermittelten Daten nicht möglich ist. Hierfür muss in technischer und organisatorischer Hinsicht sichergestellt werden, dass sowohl direkte Identifikationsmerkmale wie der Name und die genaue Adresse des

---

130 (Vgl. Wolff in Schantz/Wolff 2017, Kap. D Rn. 448)

Betroffenen als auch sonstige Hilfsmerkmale vom eigentlichen Datensatz entfernt, getrennt oder unlesbar gemacht werden.

Es müssen Richtlinien zu Gebrauch kryptographischer Maßnahmen zum Schutz der gespeicherten sensiblen Informationen entwickelt werden. Hierbei ist u.a. auch darauf zu achten, dass ein Konzept zur Lebensdauer und zum Schutz der kryptographischen Schlüssel entwickelt wird.

5.8 RECHTLICHER ASPEKT

In technischer Hinsicht ist zu prüfen, ob es sinnvoll ist, die Anonymisierung der Daten schon vor der Übertragung zum Datentreuhänder durch die erhebende Stelle durchzuführen oder eine Anonymisierung durch den Datentreuhänder vornehmen zu lassen.

In rechtlicher Hinsicht ist zu beachten, dass derzeit unterschiedliche Standards für die Anonymisierung, etwa bezüglich des Aggregationslevels von Daten, gelten. Wünschenswert wäre daher eine Harmonisierung der Interpretationen des erforderlichen Anonymisierungsstandards, z.B. durch die Datenschutzbeauftragten des Bundes und der Länder. Dies gilt umso mehr, als sich eine qualitativ hochwertige Anonymisierung von Daten angesichts der rasant schnellen, automatisierten Verarbeitung großer Datenmengen als sehr schwierig darstellt (s.o.). Wie bereits ausgeführt wurde, gibt es zahlreiche Auffassungen, nach denen eine vollständige und dauerhaft wirksame Anonymisierung nach aktuellem Stand der Technik gar nicht gewährleistet werden kann.

Insofern müssen mögliche Datenschutzrisiken fortlaufend geprüft werden. Bei solchen „Anonymisierungsstresstests" sollten zunächst Aktualität und Qualität der Anonymisierungsverfahren beurteilt werden. Stichprobenartig sollte so turnusmäßig kontrolliert werden, ob und warum die Gefahr einer De-Anonymisierung besteht.[131]

Wird die Anonymisierung durch den Datentreuhänder selbst durchgeführt, stellt die nicht anonymisierte Übertragung der Daten von der erhebenden Stelle zum Datentreuhänder eine Auftragsverarbeitung i. S. des Art. 28 EU DS-GVO dar. Es müssen also die Voraussetzungen für eine zulässige Auftragsverarbeitung, insbesondere eine Vereinbarung, aufgrund derer der Verantwortliche die Datenverarbeitung beim Auftragsverarbeiter real kontrollieren kann, vorliegen. Aus diesem Grund ist es angeraten, die Daten vor der Übermittlung zum Datentreuhänder zu pseudonymisieren oder grob zu anonymisieren, so dass die Voraussetzungen für eine Auftragsverarbeitung nicht notwendig sind. Es handelt sich dann schon nicht mehr um personenbezogene Daten, die übermittelt werden, und daher bei der Datenübertragung auch nicht um eine Auftragsverarbeitung.

Diese Vorabpseudonymisierung bzw. -anonymisierung kann allerdings bei öffentlich zugänglichen Datenbanken oder im Bereich Internet of Things nicht sichergestellt werden, da unsicher ist, ob die entsprechenden Algorithmen zur Verfügung stehen, ob auf Grund wirtschaftlicher Interessen eine solche Pseudonymisierung / Anonymisierung überhaupt von den erhebenden Stellen durchgeführt werden kann oder ob dies überhaupt gewünscht wird.

Wenn die Daten der erhebenden Stelle nicht vorab pseudonymisiert werden können, muss somit vor allem zwischen der erhebenden Stelle und dem Datentreuhänder nach Art. 28 Abs. 3 EU DS-GVO eine Vereinbarung getroffen werden, die den Gegenstand und die Dauer der Verarbeitung, die Art

---

131 (Vgl. Manske und Knobloch 2017, S. 11)

und den Zweck der Verarbeitung sowie die Art der jeweiligen personenbezogenen Daten, die Kategorien betroffener Personen sowie die Rechte und Pflichten des Verantwortlichen regeln muss. Inhalt der Vereinbarung muss insbesondere sein, dass der Datentreuhänder die personenbezogenen Daten nur auf dokumentierte Weisung des Verantwortlichen bearbeitet, dass er die Pflichten zur Datensicherheit nach Art. 32 EU DS-GVO einhält, dass ein weiterer Auftragsverarbeiter nur nach den Vorgaben des EU DS-GVO eingeschaltet wird, dass der Verantwortliche mit geeigneten technischen und organisatorischen Maßnahmen bei der Erfüllung seiner Pflichten gegenüber den Betroffenen unterstützt wird, dass nach dem Abschluss des Auftrags alle personenbezogenen Daten gelöscht oder zurückgegeben werden und dass alle mit der Verarbeitung betrauten Personen zur Vertraulichkeit verpflichtet werden. Zudem muss in der Vereinbarung gewährleistet werden, dass der Verantwortliche die erforderlichen Weisungen gemäß Art. 29 EU DS-GVO erteilen und die Auftragsvereinbarung real kontrollieren kann.

Sinnvoll erscheint im Falle einer Pseudonymisierung der Daten auch eine Trennung sensibler Informationen von Identifikationsmerkmalen in unterschiedlichen Speichern, um im Falle eines Datendiebstahls sicherstellen zu können, dass die Daten auch anonym oder hinreichend pseudonymisiert bleiben. Im Falle einer Pseudonymisierung erfolgt eine Zuordnung der beiden Datenbestände (sensible Daten auf der einen Seite, Identifikationsdaten auf der anderen Seite) über die Pseudonymisierungs-ID. Auch dies sollte vertraglich festgelegt werden bzw. Gegenstand der Vereinbarungen zwischen dem Treuhänder und der erhebenden Stelle sein.

Art. 32 Abs. 1 lit. a) EU DS-GVO fordert ausdrücklich als Handlungspflicht zur Gewährleistung des angemessenen Datensicherheitsniveaus die Pseudonymisierung und Verschlüsselung personenbezogener Daten. Auch in Erwägungsgrund 83 S. 1 EU DS-GVO wird verlangt, dass Maßnahmen zur Eindämmung von Risiken der Verarbeitung ergriffen werden, und hierzu die Verschlüsselung gezählt. Die konkreten Maßnahmen zur Gewährleistung der Datensicherheit müssen im Einzelfall bestimmt werden, um ein angemessenes Schutzniveau zu erreichen. Dieses Schutzniveau ist nicht statisch, sondern muss dynamisch im Verhältnis zwischen den bestehenden Risiken und dem Aufwand für den Schutz bestimmt werden.[132]

Somit sollten sowohl anonymisierte als auch pseudonymisierte Daten grundsätzlich verschlüsselt gespeichert werden, um bei einem Datendiebstahl das Risiko einer De-Anonymisierung oder De-Pseudonymisierung der Daten zu minimieren. Es muss die höchstmögliche Verschlüsselungsstärke gewählt werden. Regelmäßige Überprüfungen, ob die Verschlüsselungsstärke noch ausreicht, sind durchzuführen.

Die Orte, an denen die Daten gespeichert werden, müssen mit dem höchstmöglichen Sicherheitsstatus – sowohl bezogen auf die Zutrittssteuerung, als auch bezogen auf den Schutz vor umweltbedingten oder externen Bedrohungen - versehen werden. Externen Personen darf der Zugang zu diesen Anlagen in keinem Fall gestattet werden. Die Bereiche müssen durch eine Zutrittsteuerung so geschützt werden, dass lediglich authentifiziertes Personal Zutritt hat. Es wird zu einer Zwei-Faktor-Authentifizierung beim Zutritt geraten. Der Zutritt zu den Bereichen muss protokolliert werden. Diese Protokolle müssen vor unberechtigter Änderung bzw. vor Verlust geschützt werden.

---

[132] (Vgl. Wolff in Schantz/Wolff 2017, Kap. E Rn. 854 ff.) und (Vgl. Martini in Paal/Pauly 2018, Art. 32, Rn. 46)

Die Maßnahmen zum Gebrauch kryptographischer Maßnahmen und zum Gebrauch und der Lebensdauer und der Verwaltung kryptographischer Schlüssel müssen technisch umgesetzt werden.

Bei der weiteren Übermittlung der anonymisierten Daten an die abfragende bzw. auswertende Stelle hat der Treuhänder die oben unter 5.4.2 beschriebene Vorgehensweise einzuhalten, so dass eine De-Anonymisierung bei der abfragenden Stelle ausgeschlossen erscheint und es sich bei den weitergegebenen Daten nicht um personenbezogene Daten handelt.

6 FAZIT

6.1 HANDLUNGSEMPFEHLUNGEN

Zusammenfassend können die folgenden Handlungsempfehlungen abgeleitet werden:

Um eine Verhinderung von De-Anonymisierung sicherstellen zu können, ist der Aufbau eines Trust Centers notwendig, das eine Vermittlerrolle zwischen der erhebenden Stelle der Daten, den betroffenen Personen und den Akteuren, die die Daten statistisch auswerten möchten, einnimmt.

Der einzusetzende Datentreuhänder soll inhaltlich unabhängig und frei jeglicher Gewinnerzielungsabsicht sein. Eine Stellung als staatlich beliehene Stelle wird empfohlen. Die Vorteile, die der Einsatz eines Datentreuhänders mit sich bringt, sind vielfältig. So können beispielsweise Auskunftsanträge neutral geprüft und einheitlich abgewickelt werden, die Bearbeitung von Anfragen wird professionalisiert und vereinheitlicht.

Die Finanzierung des Datentreuhänders kann teilweise durch den Infrastrukturfonds erfolgen, in den die Bußgelder einfließen, die durch Nicht-Beachtung der EU DS-GVO anfallen. Eine Finanzierung sollte auch durch einen finanziellen Ausgleich durch die Anfragenden erfolgen.

Im Hinblick auf die Personalausstattung ist darauf Wert zu legen, dass die Mitarbeiter des Datentreuhänders in ihren jeweiligen Arbeitsbereichen sehr gut qualifiziert sind und regelmäßig weitergebildet werden. Jegliche Zugriffe auf die Daten müssen protokolliert werden, wobei die Protokolle vor Veränderung und Löschung geschützt sein müssen.

In technischer Hinsicht ist neben dem Einsatz hochmoderner, ausfallsicherer und zuverlässiger Hardware sehr hoher Wert auf Schutzmechanismen vor Angriffen von außen zu legen. Die Daten müssen nach den höchsten Maßstäben vor Diebstahl, aber auch vor Verlust und Zerstörung geschützt sein und verschlüsselt abgespeichert werden. Hierbei ist auf einen höchst möglichen Verschlüsselungsgrad zu achten. Es sollten regelmäßige Penetrations-Tests durchgeführt werden, um Angriffe von außen jederzeit erkennen und abwehren zu können.

Daten, die zum Datentreuhänder übertragen werden, sind vor der Übertragung von der erhebenden Stelle zu pseudonymisieren. Die Datenübertragung hat in jedem Fall verschlüsselt mit sehr hoher Verschlüsselungsstärke zu erfolgen. Beim Datentreuhänder werden die de-pseudonymisierten Daten grob anonymisiert gespeichert, wobei darauf zu achten ist, dass die Anonymisierung die Daten für zukünftige Fragestellungen nicht unbrauchbar macht. Hier ist der Einsatz von Anonymisierungsverfahren, wie k-Anonymität, l-Diversity, t-Closeness etc. nicht erforderlich.

Wird eine Forschungsfrage gestellt, ist der Datentreuhänder verpflichtet, die Daten vor der Übertragung mindestens so zu anonymisieren, dass lediglich die für die Forschungsfrage notwendigen Informationen zur anfragenden Stelle übertragen werden. Die Daten müssen beispielsweise in so großen Gruppen vorliegen, dass die in dem Kapitel 4.3.4 dargelegten Bedrohungen nicht auftreten werden. Ratsam ist, dass die Daten bei einer Anfrage durch einen Dritten durch den Datentreuhänder so anonymisiert werden, dass t-Closeness erfüllt wird. Mit Hilfe von Slicing vertauschte Werte sorgen für ein zusätzliches Maß an Sicherheit, falls die Daten beim Transfer vom Datentreuhänder zur anfragenden Stelle verloren gehen sollten. In diesem Fall ist der empfangenden Stelle mitzuteilen, in welcher

Form die Vertauschung stattgefunden hat, damit der Empfänger der Daten in die Lage versetzt wird, die ursprünglichen anonymisierten Daten wiederherzustellen.

6.2 AUSBLICK

Die vorliegende Untersuchung hat gezeigt, dass zum aktuellen Stand ein hohes Maß an Sicherheit vor De-Anonymisierung durch das Einsetzen eines Datentreuhänders zu erreichen ist. Eine 100%ige Sicherheit allerdings kann nicht garantiert werden. Dennoch ist der Einsatz eines Datentreuhänders unabdingbar, da die damit einhergehende on-demand Anonymisierung ein höchstmögliches Maß an Sicherheit vor De-Anonymisierung bringt.

Da momentan ein solcher Datentreuhänder nicht in dieser Form existiert, sollte eine solche Stelle geschaffen werden, um eine Anlaufpunkt zu haben, welcher gewährleistet, dass Analysedaten abgefragt werden können, gleichzeitig aber auch ein größtmöglicher Schutz für Daten besteht.

Im Sinne der Betroffenen sollte die Hoheit der Daten nicht in den Händen von gewinnorientierten Organisationen liegen.

Um aber die Möglichkeit einer Analyse von Daten zu schaffen, welche die Interessen der individuellen Personen und deren personenbezogene Daten schützt, gleichzeitig aber auch eine notwendige Auswertung zu ermöglichen, bedarf es des politischen Willens und Handelns, Abhilfe in diesem Zusammenhang zu schaffen.

Es liegt im fundamentalen Interesse aller Beteiligten, Organisationen, die auf Analyse von Daten angewiesen sind, eine legale und transparente, zugleich aber auch ergiebige Datenquelle bereitzustellen, welche auf die individuell anfallenden Analysetätigkeiten passende Daten liefern kann.

Die Politik ist hier gefordert, sowohl die finanziellen Möglichkeiten, als auch die weiteren Rahmenbedingungen für einen solchen Datentreuhänder zu schaffen.

Eine regelmäßige Evaluierung der Sicherheit der Daten vor unberechtigtem Zugriff und vor De-Anonymisierung muss etabliert werden, weil durch den technischen Fortschritt nicht sichergestellt werden kann, dass nach heutigem Maßstab anonyme Daten nicht später durch neue Verfahren de-anonymisiert werden können.

Der Einsatz eines Datentreuhänders scheint unter Abwägung der betrachteten Gesichtspunkte die am meisten zielführende Möglichkeit zu sein. Allerdings bedarf es weiterer Forschungstätigkeiten zu technischen Lösungen z.B. hinsichtlich kryptografischer Verfahren oder weiterer Anonymisierungstechniken, die die oben dargestellten Risiken minimieren können.

7 LITERATURVERZEICHNIS

Art. 29-Gruppe (2014). *Stellungnahme 5/2014 zu Anonymisierungstechniken, WP 216* https://docplayer.org/414072-Artikel-29-datenschutzgruppe.html. Zugegriffen: 17.06.2018

Bender, Andreas (2015). https://www.inovex.de/fileadmin/files/Fachartikel_Publikationen/Theses/anwendbarkeit-von-anonymisierungstechniken-im-bereich-big-data-andreas-bender-Mai-2015.pdf. Zugegriffen: 21.06.2018

Bode, Maximilian & Kraschewski, Daniel & Pisula, Michael & Stock, Christoph & Thüsing, Gregor (2017). *Datenschutz in Zeiten von Big Data - Moderne Methoden für gesetzeskonformen Datenschutz im Kontext von Customer Analytics und Big Data*. https://www.tngtech.com/static/user_upload/TNG_Whitepaper_Datenschutz.pdf. Zugegriffen: 28.08.2018

Burgard, Wolfram, & Stachniss, *Cyrill* (2008). *Einführung in die Informatik - Hashtables*. http://ais.informatik.uni-freiburg.de/teaching/ss08/info_MST/material/mst_14_hashing.pdf. Zugegriffen: 22.06.2018

Committee on Strategies for Responsible Sharing of Clinical Trial Data (2015). Board on Health Sciences Policy; Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. (2015). The National Academies Press.

Czernik, Agnieszka (2016). *Hashwerte und Hashfunktionen einfach erklärt*. https://www.datenschutzbeauftragter-info.de/hashwerte-und-hashfunktionen-einfach-erklaert. Fachbeitrag. Zugegriffen: 22.06.2018

de Montjoye, Yves-Alexandre & Radaelli, Laura & Singh, Vivek Kumar & Pentland, Alex (2015). *Unique in the shopping mall: On the reidentifiability of credit card metadata*. doi: 10.1126/science.1256297

Desoi, Bernd U. (2018). *Big Data und allgemein zugängliche Daten im Krisenmanagment - Exemplarische technische und normative Gestaltung von Analyse zur Entscheidungsunterstützung*. (2018). Wiesbaden: Springer Vieweg.

Dölle, Lukas (2015). *Der Schutz der Privatsphäre bei der Anfragebearbeitung in Datenbanksystemen*. https://edoc.hu-berlin.de/bitstream/handle/18452/18183/doelle.pdf?sequence=1. Zugegriffen: 25.06.2018

Dorschel, Joachim (2015). in: Dorschel, Joachim (Hrsg.). *Praxishandbuch Big Data - Wirtschaft - Recht – Technik*. (2015). Wiesbaden: Springer Gabler.

DSGVO Erwägungsgründe (o.J.). https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=DE. Zugegriffen: 15.06.2018

Dwork, Cynthia (2011). *A Firm Foundation for Private Data Analysis*. Communications of the ACM, January 2011, Vol.54 No.1. 2011. doi: 10.1145/1866739.1866758

Dwork, Cynthia, & Roth, Aaron (2014). *The Algorithmic Foundations of Differential Privacy*. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf. Zugegriffen: 29.06.2018

Eder, Johann (Interview) http://www.tmf-ev.de/News/articleType/ArticleView/articleId/1270.aspx. Zugegriffen: 29.06.2018

Els, Andrea Scripa (2017). *Artificial Intelligence as a digital Privacy Protector*. (2017). Harvard: Harvard Journal of Law and Technology, Vol. 31 No.1 2017.

Fasel Daniel (2014). *Big Data - Eine Einführung* in: *HMD* Praxis der Wirtschaftsinformatik, 08.2014, S. 386 – 400. Wiesbaden: Springer Vieweg.

Fasel, Daniel (2016). *Big Data - Grundlagen. Systeme und Nutzungspotentiale*. (2016). Fasel, Daniel & Meier, Andreas (Hrsg.). Wiesbaden: Springer Vieweg.

Fung, Benjamin C.M. & Wang, Ke & Fu, Ada Wai-Chee & Yu, Philip S. (2011). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. (2011). Boca Raton: Chapman and Hall/CRC.

Gadatsch, Andreas & Landrock, Holm (2017). Big Data für Entscheider – Entwicklung und Umsetzung datengetriebener Geschäftsmodelle. (2017). Wiesbaden: Springer Vieweg.

Gatzke, Frank (2012). *Anwendung von k-Anonymisierung auf statischen GPS-Daten* http://userpage.fu-berlin.de/semu/docs/2012_bachelor_gpsk.pdf. Zugegriffen: 29.06.2018

Ghinita, Gabriel & Karras, Panagiotis & Kalnis, Panos & Mamoulis, Nikos (2007). *Fast Data Anonymization with Low Information Loss*. in: VLDB Conference Paper. S. 758 – 769.

Goltz, Johannes (2017). *De-Anonymisierungsverfahren: Kategorisierung und deren Anwendung für Datenbankanfragen*. Bachelorthesis Rostock: Universität Rostock, Fakultät für Informatik und Elektrotechnik, Institut für Informatik, Lehrstuhl für Datenbank- und Informationssysteme. (2017).

Greveler, Ulrich, & Löhr, Dennis (2010). *Deanonymisierung von Profilen sozialer Netzwerke unter Nutzung von Metadaten aus Bildern*. P. Horster (Hrsg.) DACH Security 2010 syssec (2010) 1-10 https://www.nds.rub.de/media/nds/veroeffentlichungen/2012/07/24/Metadaten_GreLoe.PDF. Zugegriffen: 29.06.2018

Grunert, Hannes, & Heuer Andreas (2015). *Slicing in Assistenzsystemen - Wie trotz Anonymisierung von Daten wertvolle Analyseergebnisse gewonnen werden können*. in: Proceedings of the 27th GI-Workshop Grundlagen von Datenbanken, Gommern. Germany. May 26-29. 2015, S. 24 - 29.

Hauf, Dietmar (o.J.). *Allgemeine Konzepte K-Anonymity. l-Diversity and t-Closeness*. IPD Uni-Karlsruhe. https://dbis.ipd.kit.edu/img/content/SS07Hauf_kAnonym.pdf. Zugegriffen: 14.06.2018

Herbst, Tobias (2018). in Kühling, Jürgen & Buchner, Benedikt (Hrsg.). *Datenschutz-Grundverordnung/BDSG*. 2. Auflage. (2018). München: Beck C.H.

Hornung, Gerrit und Herfurth, Constantin (2018). *Datenschutz bei Big Data – Rechtliche und politische Implikationen*. in: König, Christian & Schröder, Jette & Wiegand, Erich (Hrsg.). Big Data - Chancen. Risiken. Entwicklungstendenzen. (2018). Wiesbaden: Springer VS.

IBM: *Extracting business value from the 4 V´s of big data*. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data. Zugegriffen: 18.04.2018

King, Stefanie (2014). *Big Data - Potentiale und Barrieren der Nutzung im Unternehmenskontext*. (2014). Wiesbaden: Springer.

Knopp, Michael (2015). *Pseudonym- Grauzone zwischen Anonymisierung und Personenbezug*. (2015). Wiesbaden: Springer Fachmedien.

Kohlmayer, Florian M. (2015). *Datenschutz und biomedizinische Forschung: Konzepte und Lösungen für Anonymität* (Diss.). (2015). München: Technische Universität München. Fakultät für Informatik. Lehrstuhl für medizinische Informatik.

Kühl, Eike (2015). *Sag mir, was du kaufst, und ich sag dir, wer du bist*. https://www.zeit.de/digital/datenschutz/2015-01/metadata-kreditkarten-datenschutz-anonymitaet-studie. Zugegriffen: 26.08.2018

Kurose, James, F., & Ross, Keith, W. (2012). *Computernetzwerke*. 5. Auflage. (2012). München: Pearson Studium.

LDI NRW – Landesbeauftragte für Datenschutz und Informationssicherheit NRW (2017). *Anonymität in Zeiten von Big Data.* https://www.ldi.nrw.de/mainmenu_Datenschutz/submenu_Technik/Inhalt/TechnikundOrganisation/Inhalt/Anonymitaet-in-Zeiten-von-Big-Data/Anonymitaet-in-Zeiten-von-Big-Data1.pdf, Zugegriffen: 18.06.2018

Li, Tiancheng, & Li, Ninghui & Venkatasubramanian, Suresh (2007). *t-closeness: Privacy Beyond k-Anonymitiy and l-Diversity*. doi: 10.1109/ICDE.2007.367856

Li, Tiancheng, & Li, Ninghui (2008). *Injector: Mining Background Knowledge for Data Anonymization*. Department of Computer Science. Purdue University. in: 2008 IEEE 24th International Conference on Data Engineering, S. 446 - 455.

Li, Tiancheng & Li, Ninghui & Zhang, Jian & Molloy, Ian (2009). *Slicing: A New Approach for Privacy.Preserving Data Publishing*. in: IEEE Transactions on Knowledge and Data. Engineering 24(3), S. 561 - 574.

Machanavajjhala, Aschwin & Gehrke, Johannes & Kifer, Daniel. & Venkitasubramaniam, Muthuramakrishnan (2006). *l-Diversity: Privacy Beyond k-Anonymity*, International Conference on Data Engineering, S.6ff.

Mämecke, Thorben & Passoth, Jan-Hendrik, Wehner, Josef (Hrsg.) (2018). *Bedeutende Daten - Modelle. Verfahren und Praxis der Vermessung und Verdatung im Netz*. (2018). Wiesbaden: Springer VS.

Manske, Julia & Knobloch, Tobias (2017). *Leitfaden für den Datenschutz Open Data Berlin* https://www.stiftung-nv.de/sites/default/files/policy_brief_leitfaden_open_data_datenschutz.pdf. Zugegriffen: 29.06.2018

Martini, Mario (2018). in: Paal, Boris P,, Pauly, Daniel A, (Hrsg.). *Datenschutz-Grundverordnung Bundesdatenschutzgesetz*. 2. Auflage (2018). München: C.H. Beck.

Online-Glossar des Bundesamtes für Sicherheit in der Informationstechnik https://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzKataloge/Inhalt/Glossar/glossar. Zugegriffen: 18.06.2018

o.V. Zeit-Online: *Facebook meldet weiter starke Zahlen nach Datenskandal*. https://www.zeit.de/news/2018-04/25/facebook-meldet-weiter-starke-zahlen-nach-datenskandal-180425-99-54728. Zugegriffen: 18.04.2018

o.V. Zeit-Online: *Facebook-Aktie enttäuscht beim Börsendebut*. https://www.zeit.de/wirtschaft/geldanlage/2012-05/facebook-nashdaq-handelsstart. Zugegriffen: 26.04.2018

Petrlic, Ronald, & Sorge, Christoph (2017). *Datenschutz - Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie*. (2017). Heidelberg: Springer.

Plath, Kai-Uwe & Schreiber, Lutz (2018). in: Plath, Kai-Uwe (Hrsg.). *BDSG, DSGVO: Kommentar zur DSGVO, BDSG und den Datenschutzbestimmungen des TMG und TKG*. (2018). 3. Aufl. Köln: Otto Schmidt.

Pommerening, Klaus (2005). *Pseudonyme – ein Kompromiß zwischen Anonymisierung und Personenbezug* https://www.staff.uni-mainz.de/pommeren/Artikel/pseudony.pdf. Zugegriffen: 14.06.2018

Revermann, Christoph & Sauter, Arnold (2007). Biobanken als Ressource der Humanmedizin – Bedeutung, Nutzen, Rahmenbedingungen. (2007). Berlin: Edition Sigma.

Roßnagel, Alexander & Geminn, Christian & Jandt, Silke & Richter, Philipp (2016). *Datenschutzrecht 2016 „Smart" genug für die Zukunft? - Ubiquitous Computing und Big Data als Herausforderungen des Datenschutzrechts*. (2016). Kassel: Kassel University Press.

Rust, Holger (2017). *Virtuelle Bilderwolke - Eine qualitative Big Data-Analyse der Geschmackskulturen im Internet*. Rust, Holger (2017). Wiesbaden: Springer VS.

Saft, Katrin (2017). *Erste Krankenkasse belohnt Schrittezählen*. https://www.sz-online.de/ratgeber/erste-krankenkasse-belohnt-schrittezaehlen-3587173.html. Zugegriffen: 26.08.2018

Schaar, Katrin (2016). *Was hat die Wissenschaft beim Datenschutz zukünftig zu beachten? Allgemeine und spezifische Änderungen beim Datenschutz im Wissenschaftsbereich durch die neue Datenschutzgrundverordnung*. Working Paper Series des Rates für Sozial-und Wirtschaftsdaten (RatSWD). doi: 10.17620/02671.19

Schantz, Peter & Wolff, Heinrich Amadeus (2017). *Das neue Datenschutzrecht - Datenschutz-Grundverordnung und Bundesdatenschutzgesetz in der Praxis*. München: C. H. Beck.

Schmitz, Norbert (2010). *Master Seminar Forschungsgruppe "Embedded Malware"*. Horst Görtz Institut für IT-Sicherheit. Ruhr-Universität Bochum. http://home.norbert-schmitz.de/files/Norbert_Schmitz_Deanonymisierung_paper.pdf. Zugegriffen: 16.06.2018

Schwanebeck, Axel (2017). *Gefangen im Netz - Medialer Wandel und kontinuierliche Überwachung in digitalen Welten* S. 9 – 37. In: Schröder, Michael & Schwanebeck, Axel (Hrsg.). Big Data - In den Fängen der Datenkraken: Die (un-)heimliche Macht der Algorithmen. (2017). Baden-Baden: Nomos.

Schwartmann, Rolf & Weiß, Steffen (Hrsg.). (2017). *Whitepaper zur Pseudonymisierung der Fokusgruppe Datenschutz der Plattform Sicherheit. Schutz und Vertrauen für Gesellschaft und Witschaft im Rahmen des Digital-Gipfels* https://www.gdd.de/downloads/whitepaper-zur-pseudonymisierung (2017). Zugegriffen: 16.06.2018

Statistische Ämter des Bundes und der Länder, Forschungsdatenzentren (2018). http://www.forschungsdatenzentrum.de/anonymisierung.asp. Zugegriffen: 29.06.2018

Sweeney, Latanya (2001). *Information Explosion*. in: Condentiality. Disclosure. and Data Access: Theory and Practical Applications for Statistical Agencies. (2001). https://dataprivacylab.org/dataprivacy/projects/explosion/explosion2.pdf. Zugegriffen: 29.06.2018

Sweeney, Latanya (2002). *k-anonymity: a model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10.2002, Heft 5, S. 557-570 World Scientific Publishing Company. doi: 10.1142/SO218488502001648

Thode, Jan-Christoph (2018). In: Schläger, Uwe & Thode, Jan Christoph (Hrsg.). *Handbuch Datenschutz und IT-Sicherheit*. (2018). Berlin. Heidelberg: Erich Schmidt.

Weichert, Thilo (2013). *Big Data und Datenschutz – Chancen und Risiken einer neuen Form der Datenanalyse*. in: Zeitschrift für Datenschutz. (2013). S. 251 – 259.

Zhu, Tiang & Li, Gang & Zhu, Wanlei & Yu, Philip S. (2017). *Differential Privacy and Applications* in: Advanced in Information Security 69. Heidelberg: Springer. doi: 10.1007/978-3-319-62004-6