St. Pölten University of Applied Sciences

/informatik & security

/fh/// st. pölten

Sind gängige Anonymisierungsverfahren von Bewegungsdaten im Sinne der DSGVO ausreichend?

Diplomarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

eingereicht von

Hinterholzer Matthias
is201851

im Rahmen des
Studiengangs Information Security an der Fachhochschule St. Pölten

Betreuung
Betreuer/Betreuerin: FH-Prof. Dr. Alexander Adrowitzer

Wien, 06.06.2022

(Unterschrift Autor/Autorin)

(Unterschrift Betreuer/Betreuerin)

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Diplomarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich sonst keiner unerlaubten Hilfe bedient habe.

- ich dieses Diplomarbeitsthema bisher weder im Inland noch im Ausland einem Begutachter/einer Begutachterin zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt habe.

- diese Arbeit mit der vom Begutachter/von der Begutachterin beurteilten Arbeit übereinstimmt.

Der Studierende/Absolvent räumt der FH St. Pölten das Recht ein, die Diplomarbeit für Lehre- und Forschungstätigkeiten zu verwenden und damit zu werben (z.B. bei der Projektevernissage, in Publikationen, auf der Homepage), wobei der Absolvent als Urheber zu nennen ist. Jegliche kommerzielle Verwertung/Nutzung bedarf einer weiteren Vereinbarung zwischen dem Studierenden/Absolventen und der FH St. Pölten.

Wien, 06.06.2022

(Unterschrift Autor/Autorin)

Zusammenfassung

Das digitale Zeitalter ist in vollem Gange. Standortdaten von Personen sind über soziale Medien auf schnellstem Wege ausgeforscht. Der Arbeitsplatz ist über LinkedIn durch ein bis zwei Klicks entdeckt. Das Zuhause über Standortmarkierungen auf Instagram gefunden. In der heutigen Zeit ist es keine große Herausforderung mehr, ein Bewegungsmuster für gesuchte Personen zu erstellen. Damit Personen in veröffentlichten Datensätzen nicht automatisch erkannt werden, definiert die Datenschutzgrundverordnung den Status der faktischen Anonymität. Dieser muss bei veröffentlichten europäischen Datensätzen erreicht werden und soll den Personen des Datensatzes Anonymität gewährleisten. Doch so sicher die Vorgaben der Datenschutzgrundverordnung auch klingen mögen, sie schützen die Anonymität der Personen nicht vollständig. Es existieren verschiedene Angriffsmöglichkeiten, um eine Person in einem DSGVO konformen Datensatz auch in Krisenzeiten zu identifizieren.

Diese Arbeit beschäftigt sich mit den Grundlagen des Datenschutzes. Dabei werden aktuelle und zukünftige Gefahren für die Privatsphäre des Individuums erläutert und praktische Beispiele wiedergegeben. Die drei Säulen des Datenschutzes spielen hierbei eine wichtige Rolle. Eine dieser Säulen repräsentiert die rechtlichen Regulatorien. Diese werden für ein besseres Verständnis durch eine detaillierte Beschreibung der DSGVO oberflächlich erklärt und mit den amerikanischen Datenschutzgesetzen verglichen. Des Weiteren wird beschrieben, was unter der Anonymität zu verstehen ist und vor welchen Bedrohungen sie geschützt werden muss. Das Verständnis für die unterschiedlichen Anonymisierungsverfahren ist unerlässlich, damit die technischen Maßnahmen nachvollziehbar sind. Das Ziel dieser Arbeit ist es, durch ein praktisches Beispiel die Schwächen eines gängigen Anonymisierungsverfahrens aufzuzeigen. Dazu wird ein praktisches Experiment mit realen, aus der EU stammenden Bewegungsdaten durchgeführt. Ein weiteres Ziel des Experiments liegt in der erfolgreichen Re-Identifizierung einzelner Personen in einem DSGVO-konformen anonymisierten Datensatz.

Abstract

The digital age is in full swing. People's location data is explored via social media in the fastest possible way. The workplace is discovered via LinkedIn with one or two clicks. Home found via location tags on Instagram. In today's world, it's no longer a major challenge to establish a movement pattern for people you're looking for. To prevent people from being automatically recognized in published datasets, the General Data Protection Regulation defines the status of de facto anonymity. This must be achieved in published European datasets and is intended to ensure anonymity for the individuals in the dataset. However, as secure as the requirements of the GDPR may sound, they do not fully protect the anonymity of individuals. Various means of attack exist to identify an individual in a GDPR-compliant dataset.

This thesis deals with the basics of data protection. Current and future threats to individual privacy are explained and practical examples are explained. The three pillars of data protection play an important role. One of these pillars represents the legal regulations. These are explained superficially for a better understanding by a detailed description of the GDPR and compared with the American data protection laws. Furthermore, the thesis will describe what is meant by anonymity and what threats it needs to be protected against. Understanding the different anonymization methods is essential to make the technical measures comprehensible. The aim of this thesis is to show the weaknesses of a common anonymization procedure by means of a practical example. For this purpose, a practical experiment is conducted with real transaction data originating from the EU. Another goal of the experiment is to successfully re-identify individuals in a DSGVO-compliant anonymized dataset.

Abbildungsverzeichnis

Inhaltsverzeichnis

1 Einleitung

Im ersten Kapitel dieser Diplomarbeit werden die Problemstellung sowie die Motivation diese Arbeit zu verfassen erläutert. Dabei geht es in erster Linie darum, einen ersten Blick auf die Thematik und die damit verbundenen Probleme zu werfen. Damit die Gliederung und Ziele der vorliegenden Arbeit klar abgegrenzt sind, werden diese in der Einleitung festgelegt.

1.1 Problemstellung

In den letzten Jahren hat die weltweit generierte Datenmenge zugenommen und wird sich im Jahr 2025 laut Statistik [1] auf 175 Zettabytes belaufen. Das sind umgerechnet $1{,}099\ 512\ x\ 10^{12}$ Gigabyte. Zum Vergleich, im Jahr 2018 lag diese gerade einmal bei 33 Zettabytes [1]. Durch die rasante Entwicklung neuer Softwareprodukte und die dafür benötigten Systemen, um mit der Datenflut des 21.Jahrhunderts umzugehen, wurde in der Vergangenheit der Datenschutz und die Sicherheit der Daten vernachlässigt. In Anbetracht des Umstands, dass ein Großteil der jährlich generierten digitalen Daten einen Personenbezug aufweisen, ist es ein akutes und kritisches Problem welchem entschieden entgegengewirkt werden muss. Der rapide Anstieg der produzierten Datenmenge ist unter anderem auf die Industrialisierung 4.0 [2, p. 1] zurückzuführen [3]. Diese beschreibt die übergreifende Kommunikation von Menschen und Maschinen durch intelligente und digitale Kommunikations- und Informationstechnologien. Durch die digitale Vernetzung sämtlicher Verarbeitungsstufen innerhalb der Produktionskette wird die Effizienz um ein Vielfaches gesteigert. Dabei gibt es mehrere Möglichkeiten, wie Unternehmen die Digitalisierung der Unternehmensabläufe zu ihrem Vorteil nutzen können. Ein Beispiel ist der Einsatz von Daten [3]. Hierbei werden gewonnene Daten ausgewertet und analysiert, um das Produkt in einem fortlaufenden Prozess zu optimieren. Beispielsweise sammelt Uber [4] unter anderem Geodaten seiner User, um für diese schnellstmöglich einen Fahrer zu finden. Die Kehrseite dabei ist, dass Uber sensible Daten seiner User besitzt, wie zum Beispiel die Wohnadresse oder auch das Lieblingsrestaurant. Dieser Umstand verwandelt Uber in einen großen Datenkonzern wie auch Amazon, Google oder Meta. All diese Unternehmen haben eines gemeinsam, sie verarbeiten die Daten ihrer User um ihr angebotenes Service und die User Experience fortlaufend zu verbessern. Außerdem werden die gesammelten Daten in einem aggregierten und anonymisierten Zustand an Werbetreibende weiterverkauft. [4]

Netflix stellte Videoratingdaten von ungefähr 500.000 pseudonymisierten Nutzern der Öffentlichkeit zur Verfügung. Diese enthielten Daten, wann und wie ein Netflix Nutzer einen bestimmten Film zwischen Dezember 1999 und Dezember 2005 bewertet hat. Benutzernamen wurden durch Pseudonyme ersetzt und der Datensatz veröffentlicht. Eine Gruppe von Forscher konnte nachweislich belegen, dass mit Hintergrundwissen aus der öffentlich zugänglichen IMDB Datenbank einzelne Nutzer im von Netflix Datensatz identifizierbar sind. Wie an der Identifizierung bestimmter User aus der anonymisierten Netflix Datenbank [5] zu erkennen ist, können auch anonymisierte Daten mit dem richtigen Hintergrundwissen oder durch Cross-Relationen verschiedener Datenbanken keinen ausreichenden Schutz vor Re-Identifikation bieten. Dabei ist beim ersten Blick nichts Verwerfliches an der Veröffentlichung von Film-Ratings. Wenn jedoch ein Schritt zurück gemacht wird, um das große Ganze zu überblicken, wird schnell klar, dass durch diverse Film-Ratings die politische Gesinnung oder auch individuelle Vorlieben der einzelnen User an die Öffentlichkeit gelangen. Außerdem wurden die betroffenen User im Netflix Vorfall nicht darüber informiert oder befragt, ob sie ihre Daten veröffentlichen wollen, da das veröffentlichende Unternehmen davon ausgeht, dass bei anonymisierten Daten kein Risiko besteht. Aus diesem Grund ist es wichtig, internationale Standards wie die ISO 29100 zu definieren und diese einzuhalten, damit alle Beteiligten dasselbe Verständnis von anonymisierten Daten haben. Das Ziel der Anonymisierung [6, p. 8] ist es Datenbestände so zu verändern, damit einerseits kein Personenbezug mehr hergestellt werden kann und das Individuum unerkannt bleibt, aber andererseits die Daten für weitere Analysen verwendet werden können. Bei den veröffentlichten anonymisierten Datenbeständen handelt es sich um sogenannte Mikrodaten, das sind Daten

welche personenbezogenen Informationen über Individuen enthalten. Dabei geht es zum Beispiel bei Bewegungsmustern von Taxis darum, Statistiken zu erstellen zu welchem Zeitpunkt, an welchem Wochentag, wo und wie viele Taxis an gewissen Standorten benötigt werden, um die Bedürfnisse der Kunden aber auch die wirtschaftlichen Bedürfnisse des Taxiunternehmens zu befriedigen. Mit Hilfe von Anonymisierungstechniken können diese anonymisierten Bewegungsdaten veröffentlicht und zu Forschungszwecken verwendet werden.

Neben der Industrialisierung 4.0, sind die stetig wachsenden Internetnutzer für einen großen Teil der jährlich generierten Datenmenge verantwortlich. Bei einer Bevölkerungsdichte von 7.91 Milliarden Menschen auf der Erde und einer Urbanisierung von 57 Prozent [7], sind 62.5 Prozent der gesamten Weltbevölkerung aktive Internetnutzer (Stand Jan. 2022). Mehr als die Hälfte der weltweiten Bevölkerung produziert somit Daten im Internet. Tendenz steigend. Die Anzahl der Personen welche aktiv auf sozialen Medien präsent sind, wird auf 4.62 Milliarden geschätzt, oder auch 58.4 Prozent der gesamten Weltbevölkerung [7]. Das kritische Problem daran ist, dass in den sozialen Medien vorwiegend personenbezogene Daten freiwillig und in großem Ausmaß für die Öffentlichkeit zur Verfügung gestellt werden. Dabei wird zu einem aktuellen Foto der dargestellte Aufenthaltsort verraten und somit können gewisse Bewegungsmuster der einzelnen User erstellt werden. Die daraus resultierenden Ergebnisse sind in den falschen Händen ein akutes Sicherheitsproblem für die betroffene Person.

Nicht immer ist die Bereitstellung der Daten freiwillig. Laut Allianz Risk Barometer 2022 [8, p. 10] dominieren dieses Jahr die Cyber-Risiken. Neben Ransomware Attacken führen sogenannte *data breaches* zu Imageschäden und finanziellen Verlusten. Bei einem *data breach* werden dem Unternehmen Daten gestohlen. Im schlimmsten Fall sind es personenbezogene Daten welche nicht ausreichend anonymisiert und aggregiert worden sind. Damit dieser Gefahr entgegengewirkt werden kann, müssen sich Unternehmen an gewisse Regulatorien und Standards halten. Diese dienen dem Schutz des Individuums und seinen Daten. In der EU existiert hierfür die DSGVO welche seit dem 25. Mai 2018 in Kraft [9] getreten ist und die rechtlichen Rahmenbedingungen zur Verarbeitung von personenbezogenen Daten durch Unternehmen und öffentlichen Stellen, wie in Kapitel 1 Artikel 2 der DSGVO [10, p. 1] erwähnt wird, festlegt.

Aus den aufgezählten Punkten ergibt sich die globale Herausforderung, personenbezogene Daten so weit zu anonymisieren, dass nicht nur Einzelpersonen vor Identifikation im Datenbestand geschützt sind, sondern diese Daten genauso für sinnvolle Analysen und Auswertungen verwendet werden können. Durch gezielte Forschung und der stetigen Neu- und Weiterentwicklung von Anonymisierungsmethoden wird diesem Problem entgegengewirkt.

Daraus resultiert die Forschungsfrage dieser Arbeit:

Sind gängige Anonymisierungsverfahren von Bewegungsdaten ausreichend, damit Einzelpersonen im Sinne der DSGVO selbst in Krisenzeiten wie einer Pandemie geschützt sind?

1.2 Motivation

Durch die persönliche Verwendung von Online-Vermittlungsdiensten zur Personenbeförderung und der direkten Betroffenheit durch *data breaches* sowie der Pandemie, hat das Thema rund um die Anonymisierung von personenbezogenen Daten eine hohe Präsenz in meinem Leben.

Als ich mich tiefergehend mit dem Thema beschäftig habe, merkte ich, dass ich es als interessant empfunden habe welche Schlüsse man aus Daten ziehen kann aber auch welche Gefahren in der missbräuchlichen Verwendung von Daten liegen.

Deshalb habe ich mich dazu entschieden in diesem Bereich meine Diplomarbeit zu verfassen und meinen wissenschaftlichen Beitrag zu leisten.

1.3 Methodische Vorgehensweise

Um die Forschungsfrage dieser Diplomarbeit zu beantworten, wurde ein öffentlich zugänglicher Datensatz, welcher Taximeterdaten von 442 Taxis aus Porto in Portugal zwischen dem 01.07.2013 bis zum 30.06.2014 enthält und von der Forschungswebsite „Kaggle“ [11] stammt, verwendet. Anhand dieser Daten wurde ein vertiefendes Experiment durchgeführt. Dabei wurden Taxifahrten mit gleichen Start- und Endkoordinaten individuelle Zwischenstopps hinzugefügt, um herauszufinden ab wie vielen individuellen Extrastopps ein Taxi in dem Datenbestand eindeutig identifizierbar ist. Neben der Durchführung des Experiments sind ausgewählte Fachartikel und themenspezifische Literatur in die Erstellung der vorliegenden Diplomarbeit eingearbeitet worden. Dabei sind die Quellen auf Aktualität sowie Seriosität geprüft und aussortiert, damit nur qualitativ hochwertige Informationen und bereits vorhandene Forschung als Material und Grundlage dem Literaturteil dieser Arbeit dienen. Der Literaturteil soll hierbei alle benötigten Informationen zum Verständnis des Experiments und dessen Ergebnis liefern.

Die Artikel und Informationen stammen zu einem großen Teil aus folgenden Online-Bibliotheken: Springer; Google Scholar; IT-Fachzeitschriften; IT spezifische Webseiten; Universitätsbibliotheken; Kaggle. Der verwendete Datenbestand, enthält 1 704 768 einzelne Fahrten von verschiedenen Taxis aus Porto in Portugal. Dabei enthält die Datenbank in ihrem Ursprungsformat neun Spalten. Diese enthalten verschiedene Eigenschaften der jeweiligen Taxifahrten. Das Taximeter sendet alle 15 Sekunden ein GPS-Signal, mit welchem die Fahrt rekonstruiert werden kann. Mithilfe der Software *Jupyter Notebook* wird durch schrittweises Ausführen von Programmcodestücken das Ergebnis detailliert analysiert. *Pandas* wird dazu verwendet die Daten aufzubereiten. Letzteres ist eine Datenanalyse Programmbibliothek der Programmiersprache Python. Mit dieser können Datenstrukturen bearbeitet und ausgewertet werden, um zum Beispiel Fahrten mit gleichen Start- und Endkoordinaten in einem großen Datensatz zu finden.

Die zur Analyse verwendeten Daten sind in dem Fall so weit anonymisiert, dass in erster Linie das Kennzeichen des Autos in der Datenbank fehlt bzw. durch die „Taxi_ID“ Spalte ersetzt wurde. Die vorliegende Arbeit kombiniert somit ein Experiment mit realen Bewegungsdaten sowie einen ausführlichen Literaturteil zum besseren Verständnis des Experiments.

1.4 Aufbau der Diplomarbeit

Um auf die komplexe Fragestellung der Diplomarbeit einzugehen und das Ergebnis des praktischen Experiments verständlich darlegen zu können, ist die vorliegende Arbeit in fünf Kapitel gegliedert.

Der Anfang dieser Arbeit beschäftigt sich mit den Grundlagen des Datenschutzes. Die digitalen Herausforderungen des 21. Jahrhunderts bringen einige Gefahren für die Privatsphäre des Individuums mit sich. Durch die Datensammelwut des digitalen Zeitalters werden personenbezogene Daten aller Art von Konzernen verarbeitet und weiterverkauft. Dabei ist es als Privatperson wichtig, Datenschutzbestimmungen zu lesen und sich im Klaren darüber sein, was mit den eigenen personenbezogenen Daten passiert. Der Grundbaustein für ein kritisches Bewusstsein, wird durch das Wissen über die Grundlagen des Datenschutzes und bereits vergangene Datenschutzskandale gelegt. Dabei bestehen die Säulen des Datenschutzes aus drei Komponenten. Die erste Säule besteht aus den rechtlichen Regulatorien, die vor Datenmissbrauch schützen sollen. Die zweite Säule beschäftigt sich mit Sicherheitsmaßnahmen, die ein Unternehmen zum Schutz seiner Kunden treffen kann. Die dritte und letzte Säule beschreibt die Schutzmaßnahmen, die durch die Personen selbst getroffen werden können.

Das Kapitel der *Rechtlichen Regulatorien* soll die erste Säule des Datenschutzes genauer beschreiben. Hier werden neben der detaillierten Beschreibung der DSGVO zum Vergleich die amerikanischen Datenschutzgesetze oberflächlich erklärt. Die *Allgemeinen Bestimmungen und Grundsätze* bilden die Grundlage der DSGVO. Die Rechte der betroffenen Personen werden in der DSGVO zu Gunsten der europäischen Bürger formuliert und beinhalten zum Beispiel das Recht auf Löschung. Dadurch können europäische Staatsbürger die Löschung ihrer personenbezogenen Daten bei den datenverarbeitenden Stellen anfordern. Damit die personenbezogenen Daten während und nach der Verarbeitung geschützt sind, werden die Pflichten der Datenverarbeitenden Stellen ebenfalls in der DSGVO geregelt. Dabei verpflichtet sich der/die Verantwortliche und der/die Auftragsverarbeiter*in unter Berücksichtigung des aktuellen Stands der Technik und den damit verbundenen Kosten dazu, angemessene organisatorische und technische Maßnahmen (z.B. Pseudonymisierung) zu treffen, um die faktische Anonymisierung zu gewährleisten. Diese ist laut DSGVO ausreichend, damit die anonymisierten Daten nicht mehr in den Rechtsbereich der DSGVO fallen und somit veröffentlicht werden können.

Wie die theoretischen Inhalte der vorigen Kapitel in die Praxis umgesetzt werden können, ist im vierten Kapitel festgehalten. Dabei wird erläutert was unter der Anonymität zu verstehen ist und vor welchen Bedrohungen sie geschützt werden muss. Das Verständnis für die unterschiedlichen Arten von Merkmalen ist unerlässlich, damit die technischen Maßnahmen nachvollziehbar sind. Wie die Subjekte eines Datensatzes vor der Re-Identifizierung ihrer Identität geschützt werden können, kann durch die k-Anonymität beantwortet werden. Hierbei werden Äquivalenzklassen gebildet, indem jede Wertekombination von Quasi-Identifikatoren mindestens k Personen eindeutig zugeordnet wird. Neben den *Datenaggregierenden Methoden* sind auch *Datenverändernde Methoden* in diesem Kapitel zu finden.

Damit die Forschungsfrage beantwortet wird und die Schwächen von gängigen Anonymisierungsverfahren zum Vorschein kommen, ist das Ergebnis des praktischen Experiments von großer Bedeutung. Im Beispiel werden anonymisierte GPS-Daten zu ausreichend großen Äquivalenzklassen gebildet und durch das *Rounding* Anonymisierungsverfahren eine gute Datenbasis erstellt. Das Ergebnis zeigt, sobald der Angreifer neben der Start- und Endkoordinate eine dritte GPS-Koordinate eines GPS-Protokolls weiß, kann er durch Abgleichen der dritten GPS-Koordinate mit den GPS-Protokollen der Fahrten jener Äquivalenzklasse auf ein einziges Taxi rückschließen. Damit ist die Anonymität aufgehoben und die Re-Identifizierung erfolgreich durchgeführt.

2 Grundlagen des Datenschutzes

2.1 Einführung

Die digitalen Datenströme, welche sich weltweit über den Globus erstrecken, erreichen mit der fortschreitenden Digitalisierung der Industrie und des privaten Lebensbereichs der Menschen eine neue Dimension [1]. Dabei ist die Datensammelwut nicht nur bei den bekannten Tech Giganten wie Google, Meta oder Amazon präsent. Der Sammlung von Health Daten, GPS-Daten oder auch Finanzdaten wird von den Betroffenen, welche sich in den meisten Fällen nicht bewusst sind, was mit ihren Daten geschieht, freiwillig zugestimmt. Dabei steht das ständige Optimieren seiner selbst im Vordergrund. Das reicht bis hin zur kompletten Abgabe seiner Daten zur Fremd-Überwachung durch Tech-Konzerne. Das „Life Logging" [12, p. 94] bildet hier eine Synergie mit sogenannten „Wearables" und Smartphones [13, p. 1]. Dazu zählen zum Beispiel die Apple Watch, Fitbit oder auch Garmin Uhren [14]. Im Überblick handelt es sich um Geräte, welche am Körper getragen werden und Sensoren verwenden, wie zum Beispiel GPS-Systeme, um den genauen Standort zu bestimmen [15, p. 5]. Hierbei können sensible Gesundheitsdaten wie die Herzaktivität oder die Schlafenszeit mitprotokolliert und im nächsten Schritt an ein Smartphone oder einen Laptop gesendet werden. Auf diesem werden die erhobenen Rohdaten verarbeitet und aggregiert. Dort ist die entsprechende Applikation installiert um genaue Bewegungsmuster, durchschnittliche Schlafenszeiten und zurückgelegte Distanzen des Benutzers zu berechnen [15, p. 5]. Die Aufzeichnung seiner selbst, das ständige Tracken von Fitnessdaten oder Hochladen von Fotos während sportlicher Aktivitäten erfreut sich immer größerer Beliebtheit [14]. Die Intention dahinter ist, dass aus den gewonnenen sensiblen Daten und Statistiken neue Erkenntnisse über einen selbst gewonnen werden und diese zur Optimierung verschiedener Lebensbereiche dienen.

Diese Daten besitzen nicht nur einen enormen Wert für den Benutzer, sondern auch für die Betreiber der Applikation sowie alle Drittfirmen, welche die Daten zur Verfügung gestellt bekommen. Nicht umsonst hat Apple 2014 die zentrale Gesundheitsplattform „Health" ab iOS 8 implementiert. Der Konkurrent Google stellte im Gegenzug mit Google Fit sein Produkt für Android zur Verfügung. Im Laufe der Zeit sind hier noch einige Konkurrenzprodukte hinzugekommen, wie zum Beispiel die Fitbit Uhr [12, p. 94]. Aber nicht nur physische Daten werden gesammelt. So können aktuellere Modelle mithilfe von immer besser werdender Technik, Schweißdrüsenaktivitäten und Hauttemperatur messen, um psychische Zustände zu bestimmen [15, p. 6]. Das Interesse der Konzerne an den Daten ihrer Kunden ist groß. Ebenso wichtig ist es, dass das Grundrecht auf informationelle Selbstbestimmung, welches in Österreich im Zivilrecht in den Persönlichkeitsrechten verankert ist. Es besagt, dass jede Person selbst bestimmen soll, welche Daten preisgegeben werden und wie diese weiterverarbeitet werden [15, p. 8].

2.2 Gefahren

Bei genauerer Betrachtung der Daten, werden bei sportlichen Aktivitäten neben der Aktivität selbst, in Fällen wie Runtastic oder RunKeeper zusätzlich zu den Bewegungsdaten der User, Zeitpunkte und GPS-Daten mitaufgezeichnet. Ein Problem stellen hierbei Fotos oder Gemeinschaftsläufe dar. Durch übertragene und zur Veröffentlichung gestellten Fotos kann rückgeschlossen werden, für welche Objekte sich der Läufer interessiert. Das ist für Unternehmen deshalb interessant, da diese zielgerichtete Werbung an den Läufer senden können [16, p. 3]. Durch das Taggen von Freunden oder Partnern können digitale soziale Netzwerke des Läufers kreiert und somit seine Freunde und Bekannten identifiziert werden. Ein weiteres Beispiel ist eine digitale Waage. Durch die Daten dieser ist es möglich, Beziehungen zwischen verschiedenen Menschen zu erkennen. Ebenfalls können durch Gewichtszunahmen oder Gewichtsabnahmen Rückschlüsse auf die psychische Gesundheit einer Person getroffen werden. Wenn in Kombination dazu, eine Social-Media Applikation oder auch eine Gemütszustand-Applikation verwendet wird, kann der psychische Zustand der betroffenen Person durch das Sammeln und Aggregieren seiner Präferenzen genaustens analysiert und

Werbeanzeigen maßgeschneidert zugestellt werden. Dabei gibt die betroffene Person Informationen über ihre Stimmungszustände freiwillig an [12, p. 95].

Beim *Profiling* von Einzelpersonen werden automatisiert digitale Profile der Nutzer angelegt und mit den gesammelten sensiblen Daten angereichert. Hierbei entsteht eine digitale Entität mit Vorlieben, Gesundheitswerten, einem sozialen Umfeld und psychischen Verfassung, welche durch den auswertenden Konzern entsteht. Durch die Analyse der gesammelten Daten können Verhalten, persönliche Vorlieben, Interessen, Aufenthaltsorte und Ortswechsel der getrackten Person bestimmt und möglicherweise vorhergesagt werden. [10, p. 33] So können gezielt Werbekampagnen und andere Produktplatzierung auf diversen digitalen und analogen Kanälen zugestellt werden.

Neben der kommerziellen Nutzung ist die Gefahr groß, dass die datenverarbeitenden Firmen die Daten nicht ausreichend schützen. Oftmals werden die gesammelten Daten aus der EU transferiert und in Asien oder den USA weiterverarbeitet. Xiaomi-Fitnesstracker zum Beispiel, schränkt die Rechte der Nutzer so weit ein, dass etwaige Klagen nur in China und nach chinesischem Recht erfolgen können. Der Appanbieter nutzt die Daten zu Werbezwecken und gibt dabei nicht an, ob andere sensible Gesundheitsdaten verwendet werden. Ein weiteres Beispiel ist Samsung Health. Samsung grenzt die Nutzung der Daten nicht ein und lässt dabei offen, ob weitere Daten ebenfalls genutzt werden. Die Verarbeitung erfolgt hier in Südkorea und es gibt kaum Informationen zu den Rechten von Betroffenen [14]. Außerdem sind staatliche Akteure und Geheimdienste ebenfalls an den Daten der Bürger interessiert. Der Fall Edward Snowden deckte die weltweite Überwachung der NSA auf [16, p. 3].

Damit die Verarbeitung der Daten zu Stande kommen kann, müssen die User vor dem Nutzen des Produktes die Datenschutzbestimmungen des Anbieters akzeptieren. Anderenfalls kann der Kunde das Produkt nicht nutzen. In dieser Bestimmung werden die rechtlichen Rahmenbedingungen, zur Verarbeitung der erhobenen Daten, festgelegt [12, p. 95]. Die im vorherigen Absatz genannten Beispiele sind negative Fälle, in welchen der User bei einem Datenschutzproblem das Nachsehen hat.

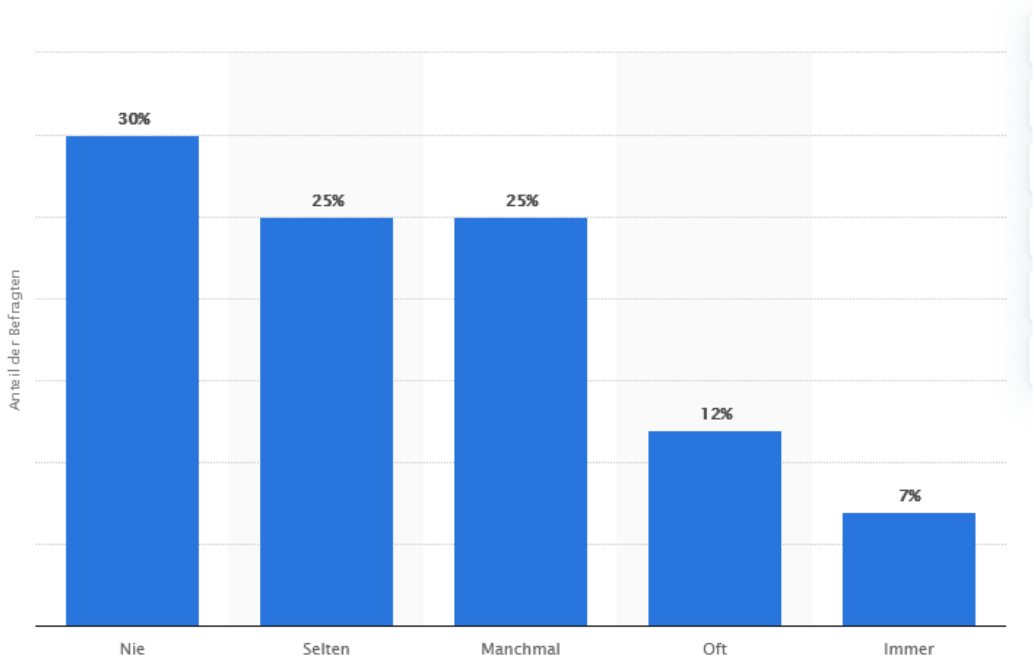

Abbildung 1: Umfrage „Lesen Sie die Datenschutzbestimmungen im Internet?“ [17]

Laut einer Statistik aus dem Jahr 2011 [17], lesen bei einer Gruppe von 1137 Personen, die über 14 Jahre alt sind, 30% davon niemals und 25% selten Datenschutzbestimmungen im Internet. Nochmal 25% geben an, dass sie sie manchmal lesen. Daraus lässt sich schließen, dass 80% der Befragten kein, bis wenig Wissen darüber besitzen wie, durch wen und wo ihre Daten verarbeitet werden. Es entsteht dahingehend eine Fremdbestimmung, dass die Konzerne die erhobenen Daten nach Belieben verarbeiten und weiterverkaufen können, solange sie sich an die gesetzlichen Regulierungen halten und die Zwecke der Verarbeitung in die Datenschutzbestimmungen schreiben. In den wenigsten Fällen müssen sie befürchten, dass jemand die Datenschutzbestimmungen tatsächlich liest [17].

Ein gutes Beispiel ist hierbei der Jö-Bonusclub in Österreich. Dieser wurde im Jahr 2021 von der Datenschutzbehörde zu einer nicht rechtskräftigen Strafe in einer Höhe von zwei Millionen verdonnert [18]. Der Grund hierfür liegt darin, dass der Jö-Bonusclub bei der Einwilligungserklärung nicht eindeutig ersichtlich angegeben hat, dass die Daten zum *Profiling* verwendet werden. Der Jö-Bonusclub hat das in erster Instanz eingesehen und dann trotz alledem die Daten von 2,3 Millionen Menschen weiterhin verarbeitet. Es wurde in der Einwilligungserklärung nicht deutlich genug hervorgehoben, was mit den Daten geschieht [18].

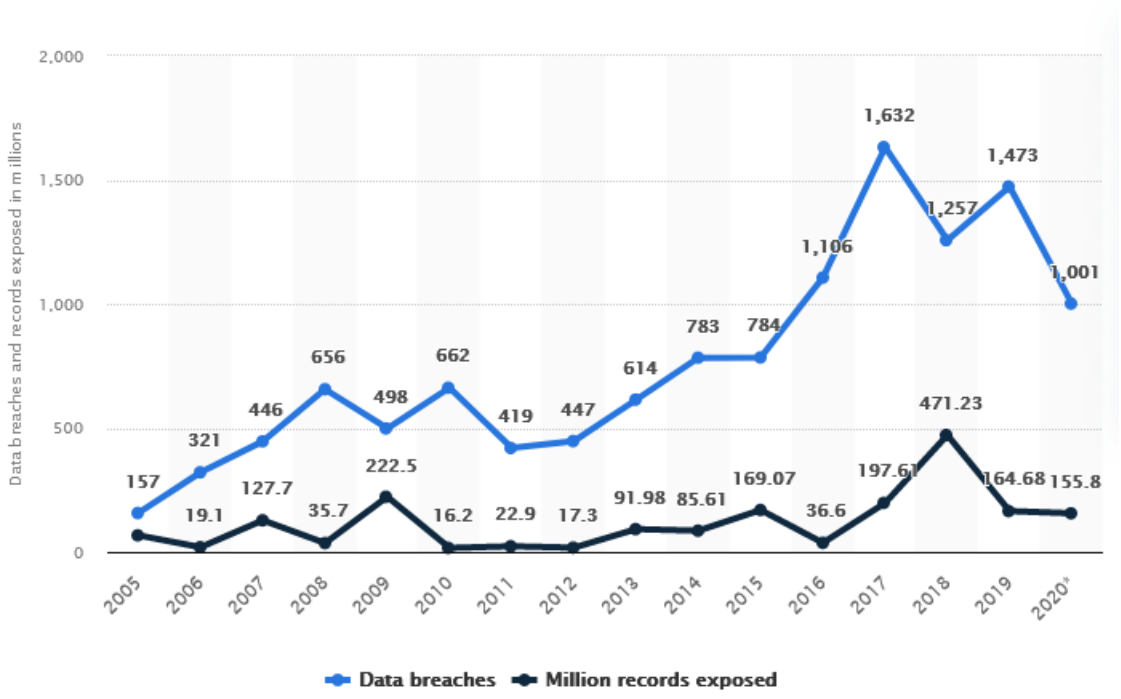

Abbildung 2: Statistik über *data breaches* und wieviele Daten geleakt wurden (USA) [19]

Eine weitere Gefahr sind kriminell motivierte Aktivitäten von Hackergruppen, welche es auf die Daten der Menschen abgesehen haben. Wie in Abbildung 2 zu sehen ist, wurden im Jahr 2020 bis zu 1001 *data breaches,* bei denen 155.8 Millionen Datensätze gestohlen wurden, gemeldet. Bei der gesamtheitlichen Betrachtung der Grafik ist zu sehen, dass jährlich mehrere Millionen Datensätze bei *data breaches* gestohlen werden. Wenn eine Person aktiv im Internet unterwegs und bei mehreren Plattformen mit ihren personenbezogenen Daten angemeldet ist, kann davon ausgegangen werden, dass diese Daten bereits öffentlich im Internet zugänglich sind [19].

2.3 Säulen des Datenschutzes

Damit die Privatsphäre der Einzelperson geschützt werden kann, ist es möglich den Datenschutz in drei Säulen zu unterteilen [16, p. 4]. Die erste Säule beschreibt die gesetzliche Regulierung durch den Staat. In Europa wird die Verarbeitung von personenbezogenen Daten sowie der Datenschutz durch die DSGVO geregelt [10, p. 1]. In Amerika gibt es diverse nationale Gesetze und in China das PIPL. Auf die genauen rechtlichen Unterschiede wird in Kapitel 3 eingegangen. Datenverarbeitende Unternehmen sind an diese rechtlichen Rahmenbedingungen gebunden und können bei Rechtsverletzungen strafrechtlich verfolgt sowie zu Strafzahlungen verurteilt werden [10, p. 24].

Die zweite Säule [20, p. 1] beschäftigt sich mit freiwilligen, erweiterbaren Sicherheitsmaßnahmen und Zertifizierungen zum Schutz der erhobenen personenbezogenen Daten durch datenverarbeitende Unternehmen. Ein Beispiel wäre hier die ISO 29100 Zertifizierung. Dieser internationale Standard definiert Anforderungen, beschreibt die Rollen im Datenverarbeitungsprozess, erklärt Datenschutzterminologie und stellt ein Framework zur Verfügung, dass der Verbesserung des Datenschutzes in Unternehmen dienen soll.

Die dritte und letzte Säule beschreibt den Selbstschutz durch den User und *Privacy Enhancing Technologies* wie zum Beispiel die Pseudonymisierung [10, p. 33] [16, pp. 13-14]. Bei der Pseudonymisierung werden die Daten so weit verschlüsselt, dass ohne Hinzuziehen von zusätzlichen Hintergrundinformationen die einzelnen Benutzer in einer Datenbank nicht mehr identifiziert werden können. Das kann durch Ersetzen des Namens mit einer ID erfolgen [10, p. 33]. Der Selbstschutz kann durch Cookie-Blocker oder anonyme Browser erzielt werden. Ein gutes Beispiel ist hier der Tor-Browser. Tor ist ein *open source* Projekt und ermöglicht das anonyme Surfen im Internet. *Open Source* bedeutet in diesem Kontext, dass der Quellcode der Applikation von jedem Interessenten eingesehen werden kann. Tor verwendet gesicherte und verschlüsselte Protokolle, um den Netzwerkverkehr seiner User zu bewerkstelligen. Dabei wird der digitale Netzwerkverkehr der User gegen böswillige Dritte geschützt und die Identität des Benutzers anonymisiert. Das führt dazu, dass es unmöglich wird die IP-Adresse zurückzuverfolgen und der User so gegen staatliche Akteure geschützt ist sowie Zensur umgehen kann [21].

2.4 Zukunft

Ein Wirtschaftszweig, der in Zukunft ebenfalls massiv von den Gesundheitsdaten seiner Kunden profitieren wird, ist die Versicherungsbranche. Dabei werden versicherten Personen vergünstigte Tarife oder Bonustarife angeboten, wenn sie ihre Gesundheitsdaten mit dem Konzern teilen. Es geht darum, eine gesunde Lebensweise zu belohnen, da für den Versicherungskonzern ein geringeres Risiko des Schadenseintritts besteht. Die Versicherten können an Frühwarnprogrammen und Schutzimpfprogrammen teilnehmen, um eine Bonusprämie zu bekommen. Hierbei variiert die Prämie in ihrer Höhe je nach Versicherungsanbieter [12, p. 96]. Ein Problem, welches sich ableiten lässt, ist die ungleiche Behandlung von verschiedenen Menschen. Chronisch Kranke zum Beispiel sind aus solchen Bonusprogrammen automatisch ausgeschlossen, da sich diese Programme nur an gesunde oder halbwegs gesunde Menschen richtet [13, p. 1]. Ein Vorreiter in der Branche ist die Generali Versicherung mit ihrer Vitality Applikation. Seit Juli 2016 können Kunden des Versicherungsdienstleisters mithilfe des Bonusprogramms und einer Anmeldung bei der Vitality App, je nach Leistung und Aktivität, Prämien und Sachleistungen erhalten [15, p. 7].

Aber nicht nur Versicherungsdienstleister sind an den Gesundheitsdaten sowie GPS-Daten seiner Nutzer interessiert. Arbeitgeber haben ebenfalls ein großes Interesse an der Gesundheit ihrer Mitarbeiter und wo sie sich aufhalten. So können Vorhersagen getroffen werden, wieviel Personal der Firma zur Verfügung steht oder in welcher Abteilung Unterstützung benötigt wird [12, p. 96]. Der erste Feldversuch dieser Art stammt aus dem Jahr 1913 durch die Ford Motor Company. Die Firma hatte nach der Einführung der Fließbandproduktion Probleme, ihre Arbeiter*innen zu halten. Daraufhin entwickelte das Unternehmen ein

Programm, in dem die teilnehmenden Mitarbeiter unter Voraussetzung eines gesunden Lebensstils doppelten Lohn verdienten. Das enthielt unangekündigte Besuche zu Hause oder Befragungen der Nachbarn. Das System stellte sich als nicht rentabel heraus und wurde eingestellt [22].

Durch die Covid-19 Pandemie sind Themen dieser Art aktueller denn je. Die Tendenz zur Überwachung der Gesundheitsdaten von einzelnen Individuen steigt stetig an. 2013, 100 Jahre später hat BP seinen Mitarbeitern Fitbit Wearables zur Verfügung gestellt, um ihre Bewegungsdaten auszuwerten und je nach Aktivität bekamen die aktiveren Nutzer Vergünstigungen und Wellness-Points [22]. Zum jetzigen Zeitpunkt handelt es sich bei diesen Beispielen um freiwillige Modelle und Arten, die Gesundheitsdaten einzelner Individuen aufzuzeichnen und zu verarbeiten. Wenn jedoch in Zukunft Personalentscheidungen oder die Gestaltung von Versicherungsprämien auf Basis der Verarbeitung von sensiblen Informationen der betroffenen Personen entschieden werden, stößt das an ethische Grenzen [12, p. 97] [22].

3 Rechtliche Regulatorien

3.1 Einführung

Wie in Kapitel 2.3 beschrieben, sind rechtliche Regulatorien eine der drei Säulen des Datenschutzes. Das Konzept zum Schutz der Privatsphäre wurde im Jahr 1948 durch die Generalversammlung der Vereinten Nationen in der „Allgemeinen Erklärung der Menschenrechte“ Artikel 12, folgendermaßen definiert:

> „Niemand darf willkürlichen Eingriffen in sein Privatleben, seine Familie, seine Wohnung und seinen Schriftverkehr oder Beeinträchtigungen seiner Ehre und seines Rufes ausgesetzt werden. Jeder hat Anspruch auf rechtlichen Schutz gegen solche Eingriffe oder Beeinträchtigungen.“ [16, p. 140]

Damals lag die Gefahr jedoch noch nicht in der digitalen Massenverarbeitung von personenbezogenen Daten, sondern in unberechtigten Wohnungsdurchsuchungen, dem Verwanzen des Telefons oder Abfangen des Briefverkehrs [16, p. 140]. Als die Computertechnik immer weiter entwickelt wurde, entstand das Bewusstsein der Menschen für die wachsende Gefahr der digitalen Datenverarbeitung und die Gefährdung ihrer Privatsphäre. Dadurch entstand der U.S. Privacy Act von 1974, welcher eine Richtlinie für die Verarbeitung von personenbezogenen Daten durch Bundesagenturen in den U.S.A. darstellt [16, p. 141] [23]. In Deutschland hingegen trat 1970, als erstes seiner Art, das Hessische Datenschutzgesetz in Kraft. Sieben Jahre später folgte mit dem Bundesdatenschutzgesetz die einheitliche Regelung des Datenschutzes für öffentliche Stellen in Deutschland. Heute agiert dieses ergänzend und unterstützt die DSGVO. Dass der Datenschutz verfassungsrechtlich stark verankert ist, zeigte das Volkszählungsurteil vom Bundesverfassungsgericht, welches einen Meilenstein in der Geschichte des Datenschutzes in Deutschland markiert [16, p. 141]. 1983 wurde die Erhebung persönlicher Daten der Bürger durch Beamte, welche der Überarbeitung des Melderegisters dienen sollte durch das Bundesverfassungsgericht untersagt, da das „informationelle Selbstbestimmungsrecht“ gefährdet war. Daraufhin wurde 1990 das Bundesdatenschutzgesetz neu gefasst [16, p. 143]. Durch die Globalisierung und den internationalen Handel sowie die immer schneller voranschreitende Digitalisierung reichten nationale Datenschutzgesetze nicht mehr aus. Damit innerhalb der EU der Austausch von personenbezogenen Daten und der freie Datenverkehr zu einem bestimmten Schutzniveau gewährleistet ist, wurde am 24.Oktober.1995 die Datenschutzrichtlinie 95/46/EG angenommen. Diese führte zur Adaptierung der nationalen Datenschutzgesetze der Mitgliedstaaten, diese sind zur Einhaltung der europäischen Richtlinien verpflichtet [16, p. 141] [24].

3.2 EU

3.2.1 Einführung

Trotz *Charta der Grundrechte* der EU von 2009, welche in Artikel 8 ein Datenschutz-Grundrecht enthält und der Datenschutzrichtlinie 95/94/EG von 1995, konnte in der EU kein adäquates Datenschutzniveau erzielt werden [16, p. 141]. Am 22.Juli 2011 folgte deshalb eine Stellungnahme des Europäischen Datenschutzbeauftragten zur Mitteilung der EU-Kommission zum Thema *Gesamtkonzept für den Datenschutz in der Europäischen Union* [24].

Dieser Stellungnahme folgte am 25.Januar 2012 ein Vorschlag der EU-Kommission zur Stärkung der digitalen Wirtschaft und der Rechte von Privatpersonen und deren Privatsphäre im Internet. Bis zur offiziellen Verordnung der DSGVO [10], wurden ein Aktionsplan zur Einführung der DSGVO [10] im EU-Raum erstellt, sowie Anpassungen und Empfehlungen der Artikel-29-Arbeitsgruppe in die Datenschutzreform eingearbeitet. Die Artikel-29-Datenschutzgruppe war ein Gremium aus unabhängigen Beratern, welche die Europäische Kommission vor Einführung der DSGVO [10] in der Thematik des Datenschutzes beriet.

Am 12.März 2014 beschloss das Europäische Parlament mit 621 Ja-Stimmen, 10 Nein-Stimmen und 22 Enthaltungen die DSGVO [10]. In der EU wurde die DSGVO [10] am 27.April 2016 als 679 Verordnung im Jahr 2016 vom Europäischen Parlament und Rat zum Schutz natürlicher Personen bei der Verarbeitung von personenbezogenen Daten und zum freien Datenverkehr beschlossen [10]. Damit wurde die Vorgänger-Richtlinie 95/46/EG (Datenschutz-Grundverordnung) aufgehoben und ersetzt. Die Verordnung ist 20 Tage nach Veröffentlichung im Amtsblatt der EU in Kraft getreten und stärkt eine Vielzahl der bereits bestehenden Rechte von Individuen. Das Recht auf Löschung, dass ein Unternehmen die personenbezogenen Daten löscht, wenn diese für die Verarbeitung nicht mehr benötigt werden oder die Einwilligung zur Verarbeitung revidiert wird, ist eine der vielen besonderen Errungenschaften [10, p. 1] [24].

Ab dem 25.Mai 2018 wurde die DSGVO [10] in der EU eingeführt und fand ab diesem Zeitpunkt Anwendung. Einige Unternehmen, zum Beispiel Unternehmen deren Hauptaufgabe die Überwachung oder Verarbeitung von personenbezogenen Daten ist, sowie Unternehmen des öffentlichen Bereichs mussten ab diesem Zeitpunkt einen Datenschutzbeauftragten ernennen. Dieser war dafür zuständig, dass die DSGVO [10] im Unternehmen eingehalten wird [24].

Die durch die DSGVO [10] geschaffenen Rahmenbedingungen zur Verarbeitung von personenbezogenen Daten, beinhalten Grundpfeiler, welche die Grenzen und den rechtlichen Spielraum der Unternehmen eingrenzen. Außerdem werden die betroffenen Personen mit erhöhten Rechten ausgestattet, um bis zu einem gewissen Grad entscheiden zu können, wie ihre Daten verarbeitet werden dürfen und wann sie zu löschen sind [25].

In den nachfolgenden Absätzen wird die Person, von der die personenbezogenen Daten verarbeitet werden als *betroffene Person* bezeichnet. Die datenverarbeitende Stelle wird als *die/der Verantwortliche* bezeichnet.

3.2.2 Allgemeine Bestimmungen und Grundsätze der DSGVO [10]

Kapitel 1, **Artikel 1 bis Artikel 4** der DSGVO [10] beschäftigen sich mit den allgemeinen Bestimmungen. Der erste Artikel dient hauptsächlich der Erläuterung und Eingrenzung der Datenschutzgrundverordnung. Die DSGVO [10] dient sowohl dem Schutz von natürlichen Personen bei der Datenverarbeitung als auch dem Schutz des freien Verkehrs solcher Daten [10, p. 2] [26]. **Artikel 2 und 3** behandeln die sachlichen und räumlichen Anwendungsbereiche der DSGVO [10]. Als sachlichen Anwendungsbereich wird die ganze oder teilweise automatisierte Datenverarbeitung von personenbezogenen Daten verstanden. Daten, die nicht automatisiert verarbeitet werden, fallen nur dann in den Rechtsbereich der DSGVO [10], wenn die Speicherung der Daten in einem Dateisystem vorgesehen ist. Ausgenommen sind Fälle der Datenverarbeitung in denen kleine bis mittlere Unternehmen, falls sie als natürliche Person organisiert sind und nur im persönlichen Umfeld Daten verarbeiten. Weiters ist die Verarbeitung von Daten, welche die nationale und internationale Sicherheit der europäischen Mitgliedstaaten betrifft, ebenfalls ausgenommen. Dazu zählt sowohl die Strafverfolgung sowie die Bereiche der öffentlichen Sicherheit [10, pp. 2-3] [26]. Der räumliche rechtliche Rahmen wird durch Artikel 3 festgelegt. Dabei wird mit rechtlichen Unsicherheiten aufgeräumt und folgende datenverarbeitende Stellen zur Einhaltung der DSGVO [10] gezwungen: [10, p. 3] [27]

- Stellen, die personenbezogene Daten im Rahmen der Tätigkeit einer Niederlassung von Verantwortlichen oder Auftragsverarbeitenden in der europäischen Union, verarbeiten, unabhängig davon, ob die Datenverarbeitung in der Union geschieht [10, p. 3] [27].
- Stellen, die personenbezogene Daten durch eine Verkaufstätigkeit in der europäischen Union verarbeiten [10, p. 3] [27].
- Stellen, die personenbezogenes Verhalten von Personen beobachten, sofern das Verhalten in der Union stattfindet [10, p. 3] [27].

- Stellen, die personenbezogene Daten außerhalb der EU verarbeiten, die aber aufgrund völkerrechtlicher Bestimmungen der Europäischen Union zuzurechnen sind [10, p. 3] [27].

Datenverarbeitende Stellen müssen durch Artikel 2 und 3 bestimmen, ob sie in den Geltungsbereich der DSGVO [10] fallen [10, p. 3] [27].

Artikel 4 beinhaltet Begriffsbestimmungen zum besseren Verständnis der DSGVO [10]. Die Wichtigsten sind wie folgt: [28] [10, pp. 3-4]

- „personenbezogene Daten“ sind alle Daten, durch die eine einzelne natürliche Person eindeutig identifizierbar ist. Darunter fallen zum Beispiel der Name, Standortdaten oder besondere Merkmale einer Person über diese sie in einer beliebig großen Menge an unbekannten Personen ausgemacht werden kann [10, p. 3] [28].
- Unter „Verarbeitung“ fällt jede durchgeführte Datenverarbeitung und alle dazugehörenden Vorgänge angefangen vom Erheben, bis hin zur Vernichtung der Daten [10, p. 3].
- „Profiling“ siehe Kapitel 2.2.
- „Pseudonymisierung“ siehe Kapitel 2.3.
- Unter „Dateisystem“ wird jede strukturierte Sammlung von kategorisierten personenbezogenen Daten verstanden [10, p. 4] [28].

Kapitel 2, **Artikel 5 bis 11** [10, pp. 6-7] [29] beschreiben in umfassender Form, die Grundsätze der Verarbeitung von personenbezogenen Daten in der Europäischen Union. Artikel 5 stellt hierbei die Grundpfeiler für die Verarbeitung der Daten vor. Diese besagen, dass die Verarbeitung der personenbezogenen Daten rechtmäßig, transparent und unter Beachtung von Treu und Glauben zu verarbeiten sind. Weiters ist das Prinzip der Datenminimierung zu verfolgen. Das bedeutet, dass Daten nur für festgelegte und legitime Zwecke erhoben werden dürfen. Der Umfang der erhobenen Daten, darf hierbei das angemessene Maß nicht überschreiten. Die Speicherung der Daten darf nur für den benötigten Zeitraum der Verarbeitung und zur Erfüllung des Zwecks der Erhebung erfolgen. Die Grundsätze der Vertraulichkeit und Integrität sind ebenfalls zu beachten. Das bedeutet es sind angemessene organisatorische sowie technische Maßnahmen zu treffen, um diese Grundsätze zu erfüllen [10, pp. 6-7] [29].

Der **6 Artikel** definiert mögliche Bedingungen für eine rechtmäßige Verarbeitung von personenbezogenen Daten. Darunter fallen: [10, pp. 7-8] [30]

- Die/Der Betroffene unterzeichnet eine elektronische oder schriftliche Einwilligung [10, pp. 7-8] [30].
- Die Datenverarbeitung ist zur Erfüllung einer rechtlichen Verpflichtung, dessen Vertragspartei die/der Betroffene ist, erforderlich [10, pp. 7-8] [30].
- Die Verarbeitung, ist zum Schutz lebenswichtiger Interessen der betroffenen Person oder einer anderen natürlichen Person, notwendig [10, pp. 7-8] [30].
- Die Verarbeitung ist für das öffentliche Interesse oder die Ausübung öffentlicher Gewalt notwendig [10, pp. 7-8] [30].

Weiters werden in **Artikel 6** Absatz 4, Fälle in denen sich die Verarbeitungszwecke während der Verarbeitung ändern, behandelt. Wenn in diesen Fällen keine weitere Einwilligung seitens der Betroffenen besteht, muss durch die/den Verantwortliche*n geprüft werden, ob der neue Zweck mit dem Zweck der ursprünglichen Erhebung übereinstimmt. Dabei müssen vor allem die möglichen Folgen, die durch die Weiterverarbeitung für die betroffenen Personen eintreten, evaluiert werden [10, pp. 8-9] [30].

Artikel 7 stellt Bedingungen für die Einwilligung auf. Absatz 1 verpflichtet die/den Verantwortliche*n, im Falle, dass die Datenverarbeitung auf einer Einwilligung der betroffenen Person besteht, diese nachweisen zu können. In den Absätzen 2 und 3 wird von der Einwilligung erwartet: [10, p. 9] [31]

- Dass der betroffenen Person in einer klaren und einfachen Sprache kommuniziert wird, welchen Verarbeitungszwecken und Vorgängen sie zustimmt [10, p. 9] [31].
- Dass die betroffene Person über ihr Widerrufsrecht informiert wird [10, p. 9] [31].
- Dass die Beurteilung, ob die Einwilligung freiwillig erfolgte und die Erfüllung des Vertrags von einer Verarbeitung von personenbezogenen Daten abhängig ist, die für die Erfüllung des abgeschlossenen Vertrags nicht zwingend notwendig ist, erfolgt [10, p. 9] [31].

Artikel 8 [10, pp. 9-10] handelt vom Jugendschutz und schützt alle betroffenen Personen, die das 16te Lebensjahr noch nicht vollendet haben. Personenbezogenen Daten von Kindern die jünger als die festgelegte Altersgrenze sind, können ausschließlich durch die Zustimmung ihrer erziehungsberechtigten Personen rechtmäßig verarbeitet werden. Die Mitgliedstaaten sind berechtigt, die erforderliche Altersgrenze bis auf das 13te vollendete Lebensjahr zu senken.

Artikel 9 [10, pp. 10-11] beschäftigt sich mit der Verarbeitung von personenbezogenen Daten besonderer Kategorien. Dabei dürfen sensible Merkmale, wie zum Beispiel die rassische oder ethnische Herkunft sowie die politische Gesinnung einer Person nicht verarbeitet werden, falls diese nicht ausdrücklich zustimmt. Weitere Ausnahmen sind Datenverarbeitungsvorgänge die:

- dem Schutz von lebenswichtigen Interessen von betroffenen Personen dient [10, pp. 10-11].
- sich auf die Verarbeitung von bereits von der betroffenen Person veröffentlichten personenbezogenen Daten bezieht [10, pp. 10-11].
- zu Archivzwecken durchgeführt werden [10, pp. 10-11].
- im Zusammenhang von Gesundheitsgefahren verarbeitet werden [10, pp. 10-11].

Artikel 10 [10, pp. 11-12] regelt die Sicherheitsvorkehrungen in Bezug auf die Verarbeitung von strafrechtlich relevanten personenbezogenen Daten. Die rechtmäßige Verarbeitung dieser speziellen Daten darf ausschließlich unter Einhaltung von Sicherungsvorkehrungen und nur unter behördlicher Aufsicht erfolgen. Artikel 11 sieht vor, dass falls die Identifizierung der betroffenen Person durch die/den Verantwortliche*n für die Zwecke der Datenverarbeitung nicht erforderlich ist, dieser zur bloßen DSGVO-Konformität keine zusätzlichen Informationen zur Identifikation der betroffenen Person aufzubewahren hat [10, pp. 11-12].

3.2.3 Rechte der betroffenen Personen

Damit die Rechte der betroffenen Personen offen kommuniziert werden, wird in Kapitel 3, Abschnitt 1, **Artikel 12** in der DSGVO [10, p. 12] die transparente Kommunikation der Bedingungen für die Verarbeitung der personenbezogenen Daten der betroffenen Person gefordert. Die datenverarbeitende Seite muss geeignete Maßnahmen treffen, um der Person in akkurater, transparenter und verständlicher Form ihre Rechte mitzuteilen. Besonders wenn es um die Verarbeitung personenbezogener Daten von Minderjährigen geht. Die Zustellung der Modalitäten muss in schriftlicher oder elektronischer Form erfolgen. Falls von der betroffenen Person gefordert, kann die Kommunikation der Modalitäten auch in mündlicher Form erfolgen. Hierbei muss jedoch die Identität der betroffenen Person nachgewiesen werden können [10, p. 12]. Nach Anfrage der betroffenen Person, muss die verarbeitende Stelle innerhalb eines Monats, Informationen zu den nach Artikel 15-22 ergriffenen Maßnahmen zur Verfügung stellen. Falls die Komplexität und die Anzahl der Anfragen es erfordern, kann die gesetzte Frist um einen Monat verlängert werden, damit die verantwortliche Stelle genügend Zeit für eine aufklärende Antwort hat. Der/Die Verantwortliche muss nach der ersten Frist die betroffene Person bezüglich der Fristverlängerung informieren und diese begründen. Hat der/die Verantwortliche nachvollziehbare Zweifel an der Anfrage, so kann er von der anfragenden Person

weitere Informationen verlangen, die zur Identifizierung der anfragenden Person beitragen. Mithilfe dieser zusätzlichen Informationen sollte es der verantwortlichen Stelle möglich sein die begründeten Zweifel aus dem Weg zu räumen [10, p. 13].

Abschnitt 2, **Artikel 13** [10, p. 14] [25] beschäftigt sich mit der Informationspflicht bei der Erhebung von personenbezogenen Daten bei der betroffenen Person. Zum Zeitpunkt der Erhebung der personenbezogenen Daten, muss der/die Verantwortliche, der betroffenen Person den Namen sowie die Kontaktdaten des/der Verantwortlichen oder seines/ihres Vertreters mitteilen. Falls ein/e Datenschutzbeauftragte*r bei der datenverarbeitenden Stelle tätig ist, müssen seine Kontaktdaten der betroffenen Person kommuniziert werden. Dieser Vorgang dient der leichteren Kommunikation zwischen Betroffenen und Verantwortlichen. Für das bessere Verständnis der Betroffenen müssen die Gründe, für die die personenbezogenen Daten verarbeitet werden, sowie die Rechtsgrundlage bekannt sein. Darüber hinaus müssen die Kategorien der verarbeitenden Daten, sowie die Empfänger oder Kategorien von Empfänger offengelegt werden. Darauf folgt eine für die Betroffenen nachvollziehbare sowie transparente Verarbeitung ihrer Daten. Falls personenbezogene Daten in ein Drittland oder eine internationale Organisation übermittelt werden, ist das der betroffenen Person mitzuteilen [10, p. 14] [25].

Damit eine transparente und faire Verarbeitung der personenbezogenen Daten [10, p. 14] [25] möglich ist, sind zusätzlich zu den genannten Punkten die Dauer, für die die Daten gespeichert werden bekanntzugeben. Wenn dies nicht möglich ist, sind die Kriterien zur Bestimmung dieser Dauer offen zu legen. Außerdem muss die betroffene Person über ihr bestehendes Recht der Auskunft, Berichtigung oder Löschung ihrer Daten, Einschränkung der Verarbeitung oder eines Rechts auf Widerspruch gegen die Verarbeitung sowie das Recht auf Datenübertragbarkeit informiert werden. Die betroffene Person muss zusätzlich darüber informiert werden, dass sie ein Beschwerderecht bei der Aufsichtsbehörde besitzt. Außerdem muss ihr bekannt sein, ob die Bereitstellung der personenbezogenen Daten vertraglich, gesetzlich oder für einen Vertragsabschluss notwendig ist. Die Folgen einer Nichtbereitstellung der Daten muss ebenfalls festgehalten werden. Falls die personenbezogenen Daten für einen anderen Zweck weiterverarbeitet werden, für den sie erhoben wurden, so muss die/der Verantwortliche*r der betroffenen Person Informationen über den Zweck sowie alle maßgeblichen Informationen über die Weiterverarbeitung zur Verfügung stellen [10, p. 14] [25].

Abschnitt 2, **Artikel 15** [10, pp. 16-17] [25] definiert die Auskunftsrechte der betroffenen Person. Sie/Er hat das Recht darüber informiert zu werden, ob und warum ihre/seine personenbezogenen Daten verarbeitet werden. Darüber hinaus müssen im Falle der automatisieren Entscheidungsfindung einschließlich *Profiling*, aussagekräftige Informationen über die eingesetzte Logik sowie die Auswirkungen dieser Art von Datenverarbeitung der betroffenen Person mitgeteilt werden [10, pp. 16-17] [25].

Abschnitt 3, **Artikel 16 und 17** [10, p. 17] [25] definieren das Recht auf Berichtigung und Löschung der verarbeitenden Daten. Die betroffenen Personen sind in der Lage, von der/dem Verantwortlichen die Berichtigung oder Ergänzung von nicht vollständigen oder falschen personenbezogenen Daten zu verlangen. Das Recht auf Löschung bzw. das *Recht auf Vergessenwerden* ermöglicht es der betroffenen Person die unverzügliche Löschung ihrer personenbezogenen Daten unter dem Vorwand folgender Gründe zu verlangen [10, p. 17] [25].

- Falls die personenbezogenen Daten für die Zwecke, für die sie erhoben wurden, nicht mehr notwendig sind [10, p. 17] [25].
- Die betroffene Person ihre Einwilligung zur Verarbeitung widerruft und es an einer anderen Rechtsgrundlage zur Verarbeitung fehlt [10, p. 17] [25].
- Die personenbezogenen Daten unrechtmäßig verarbeitet wurden oder ihre Löschung zur Erfüllung rechtlicher Verpflichtungen durchgeführt werden muss [10, p. 17] [25].

Hat die/der Verantwortliche die personenbezogenen Daten veröffentlicht und ist zu deren Löschung verpflichtet, so muss er/sie mithilfe verfügbarer Technologien und der damit zusammenhängenden Kosten angemessene Maßnahmen treffen, um alle datenverarbeitenden Verantwortlichen, die die veröffentlichten, personenbezogenen Daten verarbeiten, darüber zu informieren, dass eine betroffene Person die Löschung aller Verlinkungen zu diesen Daten oder Kopien dieser personenbezogenen Daten verlangt hat [10, p. 18].

Die im vorherigen Absatz genannten Gründe sind in folgenden Fällen ungültig:

- Falls die Verarbeitung der Daten zur Ausübung des Rechts auf freie Meinungsäußerung notwendig ist [10, p. 18].
- Die Verarbeitung zur Erfüllung rechtlicher Verpflichtungen nach dem Recht der Europäischen Union oder der ihrer Mitgliedstaaten, dem der Verantwortliche unterliegt, erfordert [10, p. 18].
- Die Verarbeitungsgründe im öffentlichen Interesse im Bereich der öffentlichen Gesundheit liegen [10, p. 18].
- Zur Geltendmachung, Ausübung oder Verteidigung von Rechtsansprüchen [10, p. 18].

In **Artikel 18** [10, pp. 18-19] [25] wird das Recht auf Einschränkung der Verarbeitung von personenbezogenen Daten geregelt. Wenn die Richtigkeit der personenbezogenen Daten von der betroffenen Person bestritten wird, kann sie die Einschränkung der Verarbeitung ihrer Daten fordern. Während dieser Dauer hat der/die Verantwortliche die Möglichkeit die Richtigkeit der personenbezogenen Daten zu prüfen. Ein weiterer Grund zur Einschränkung der Verarbeitung ist, falls die Verarbeitung der personenbezogenen Daten unrechtmäßig stattfindet, die betroffene Person die Löschung dieser Daten ablehnt und stattdessen die Einschränkung der Nutzung der Daten fordert. Darüber hinaus kann die Verarbeitung eingeschränkt werden, falls der/die Verantwortliche die personenbezogenen Daten für die Zwecke der Verarbeitung nicht mehr benötigt, die betroffene Person diese Daten jedoch zur Ausübung, Verteidigung oder Geltendmachung von Rechtsansprüchen benötigt. Wurde die Verarbeitung der personenbezogenen Daten eingeschränkt, dürfen die Daten nur mit Einwilligung der betroffenen Person, zur Geltendmachung diverser Rechtsansprüche oder aus Gründen eines wichtigen öffentlichen Interesses der Europäischen Union oder eines Mitgliedstaates verarbeitet werden. Falls die Einschränkung stattfindet, muss der/die Verantwortliche der betroffenen Person dies mitteilen [10, pp. 18-19] [25].

Abschnitt 3, **Artikel 19** [10, p. 19] verpflichtet die Verantwortlichen, allen Empfängern denen personenbezogene Daten preisgegeben wurden, jede Löschung oder Berichtigung der personenbezogenen Daten oder eine Einschränkung ihrer Verarbeitung mitzuteilen. Es sei denn dieser Vorgang ist unmöglich oder mit einem unverhältnismäßigen Aufwand verbunden [10, p. 19].

Artikel 20 [10, pp. 19-20] definiert das Recht auf Datenübertragbarkeit. Dies besagt, dass die betroffene Person das Recht hat, ihre personenbezogenen Daten, die sie der/dem Verantwortlichen zur Verfügung gestellt hat, in einem strukturierten und maschinenlesbaren Format zu erhalten. Sofern die Verarbeitung auf einer Einwilligung beruht oder die Verarbeitung mithilfe von automatisierten Verfahren erfolgt, darf die betroffene Person ihre strukturierten Daten einer anderen datenverarbeitenden Stelle, ohne Behinderung durch die erste verarbeitende Stelle, übermitteln. Die betroffene Person hat das Recht, sofern dies technisch möglich ist, dass die personenbezogenen Daten direkt von einer/m Verantwortlichen an die nächste verantwortliche Stelle übermittelt werden. Dieses Recht darf die Freiheiten und Rechte anderer Personen nicht beeinträchtigen. Dieses Recht ist für eine Datenverarbeitung, die für die Ausführung einer Aufgabe im öffentlichen Interesse liegt, ungültig [10, pp. 19-20].

Abschnitt 4, **Artikel 21** [10, pp. 19-20] ermöglicht der betroffenen Person aus Gründen, die sich aus einer bestimmten Situation ergeben, zu jeder Zeit gegen die Verarbeitung ihrer personenbezogenen Daten Widerspruch einzulegen. Dieser Widerspruch gilt ebenso für das *Profiling*. Die/Der Verantwortliche ist daraufhin verpflichtet die Verarbeitung der personenbezogenen Daten einzustellen, es sei denn, er kann

dringende schutzwürdige Gründe nennen, die die Rechte, Freiheiten und Interessen der betroffenen Person überwiegen. Außer die Verarbeitung dient der Ausübung, Verteidigung oder Geltendmachung von Rechtsansprüchen. Falls die Verarbeitung von personenbezogenen Daten der Direktwerbung dient, so kann die betroffene Person jederzeit Widerspruch gegen die Verarbeitung ihrer Daten, zum Zwecke dieserart von Werbung, einlegen. Die betroffene Person muss zum Zeitpunkt der ersten Kontaktaufnahme auf das Recht auf Widerspruch hingewiesen werden. Dieser Hinweis muss gesondert von anderen Informationen sein und dabei in leicht verständlicher Form erfolgen. Der Widerspruch der betroffenen Person kann automatisiert mittels technischer Hilfsmittel erfolgen [10, p. 20].

In **Artikel 23** [10, pp. 21-22] steht geschrieben, dass die Rechte der betroffenen Personen durch Rechtsvorschriften der Europäischen Union oder ihrer Mitgliedstaaten, denen die/der Verantwortliche unterliegt, im Wege von Gesetzgebungsmaßnahmen beschränkt werden können, sofern diese Beschränkung die Grundfreiheiten achtet und die in einer demokratischen Gesellschaft vorherrschenden Werte nicht verletzt bzw. eine verhältnismäßige Maßnahme darstellt, die Folgendes sicherstellt: [10, pp. 21-22]

- Die nationale, öffentliche Sicherheit oder Landesverteidigung [10, pp. 21-22].
- Die Verfolgung und Strafvollstreckung, einschließlich des Schutzes von Gefahren für die öffentliche Sicherheit [10, pp. 21-22].
- Den Schutz von wichtigen finanziellen oder wirtschaftlichen Interessen der Europäische Union oder eines Mitgliedstaates. Besonders im Haushalts-, Währungs-, und Steuerbereich sowie in den Bereichen der sozialen Sicherheit und der öffentlichen Gesundheit [10, pp. 21-22].
- Den Schutz der Unabhängigkeit von Justiz und den Schutz der Gerichtsverfahren [10, pp. 21-22].
- Damit Kontroll-, Überwachungs- und Ordnungsfunktionen, die durch Ausübung von öffentlicher Gewalt zum Zwecke der vorherigen Punkte, möglich sind [10, pp. 21-22].
- Den Schutz der betroffenen Person oder die Rechte und Freiheiten anderer Personen [10, pp. 21-22].
- Die Vollstreckung zivilrechtlicher Ansprüche [10, pp. 21-22].

Alle Gesetzgebungsmaßnahmen die im Sinne von **Artikel 23** getroffen werden, müssen spezifische Vorschriften enthalten, zumindest in Bezug auf: [10, p. 22]

- Die Zwecke der Verarbeitung oder die Verarbeitungskategorien, sowie die Kategorien personenbezogener Daten [10, p. 22].
- Die Garantien gegen Missbrauch oder nicht rechtmäßiger Übermittlung sowie nicht rechtmäßigem Zugang [10, p. 22].
- Den Umfang der in der Gesetzgebungsmaßnahme getroffenen Beschränkung [10, p. 22].
- Die Kontaktdaten der Verantwortlichen [10, p. 22].
- Die entsprechende Speicherfristen sowie die geltenden Garantien unter der Berücksichtigung von Umfang, Zwecken und Art der Verarbeitung [10, p. 22].
- Die Risiken für die Freiheiten und Rechte der betroffenen Personen sowie das Recht auf Informationen über die Beschränkung, sofern dies nicht den Grund der Beschränkung widerspricht [10, p. 22].

3.2.4 Pflichten der Verantwortlichen und Auftragsverarbeiter

Laut der DSGVO [10] wurde die Bezeichnung *datenschutzrechtlicher Auftraggeber* auf „die/der Verantwortliche“ geändert. Darunter fallen natürliche oder juristische Personen, Einrichtungen, Behörden oder andere Stellen, welche über die Mittel und Zwecke der Verarbeitung von personenbezogenen Daten bestimmen [32]. Der Verantwortliche hat laut der DSGVO [10] bestimmten Pflichten nachzukommen. Eine dieser Pflichten ist die Informationspflicht gegenüber betroffenen Personen. Hierbei ist der Verantwortliche

verpflichtet sämtliche Anfragen und Anträge von betroffenen Personen im Sinne der in Kapitel 3.2.1 erwähnten Betroffenenrechte, zu bearbeiten [32].

In Kapitel 4, Abschnitt 1, **Artikel 24-25** [10, p. 23] [33] werden die allgemeinen Pflichten und Verantwortungsbereiche des Verantwortlichen in Bezug auf den Datenschutz durch Technik definiert [10, pp. 22-23]. Dabei verpflichtet sich der/die Verantwortliche und der/die Auftragsverarbeiter*in unter Berücksichtigung des aktuellen Stands der Technik und den damit verbundenen Kosten dazu, angemessene organisatorische und technische Maßnahmen wie zum Beispiel die Pseudonymisierung zu treffen, um ein adäquates Schutzniveau zu ermöglichen. Die Umstände, der Umfang und die Zwecke der Verarbeitung müssen analysiert und dabei die unterschiedlichen Eintrittswahrscheinlichkeiten sowie die Schwere des Risikos für die Freiheiten und Rechte natürlicher Personen eruiert werden. Sofern die zu treffenden Maßnahmen im Verhältnis zur Verarbeitung der personenbezogenen Daten stehen, müssen diese geeignete Datenschutzvorkehrungen durch die Verantwortlichen getroffen werden. Die in Artikel 42 festgehaltenen Zertifizierungsregeln und Standards können herangezogen werden, um die Pflichten der Verantwortlichen nachzuweisen [10, p. 23]. In Artikel 25 ist außerdem geschrieben, dass der/die Verantwortliche sicherstellen muss, dass durch technische Voreinstellungen nur jene personenbezogenen Daten verarbeitet werden, welche für den vorab definierten Verarbeitungszweck notwendig sind. Diese Maßnahme zum Datenschutz gilt für den vollen Umfang der Verarbeitung, ihrer Zugänglichkeit, ihrer Speicherfrist und für die Menge an erhobenen personenbezogenen Daten [10, p. 23] [33].

Falls mehr als eine Person als Verantwortliche/r definiert wird, so legen sie gemeinsam die Mittel und Zwecke der Verarbeitung fest. Dadurch sind sie als Gruppe für die Verarbeitung der personenbezogenen Daten verantwortlich. Die betroffenen Personen können ihre Rechte im Zuge ihrer Betroffenenrechte gegenüber jeden einzelnen Verantwortlichen durchsetzen [10, pp. 23-24] [32]. Verantwortliche, die Ihren Sitz außerhalb der Europäischen Union haben, die sich aber trotzdem im Rechtsbereich der DSGVO [10] befinden, müssen in schriftlicher Form einen Vertreter ernennen. Es ist zwingend notwendig, dass der Vertreter in einem der Mitgliedstaaten eine Niederlassung betreibt, in denen die personenbezogenen Daten der betroffenen Personen im Zuge der verkauften Waren oder Dienstleistungen erhoben werden. Die Niederlassung des Vertreters fungiert somit zusätzlich oder anstelle des Verantwortlichen als Auskunftsstelle für betroffene Personen oder Aufsichtsbehörden [32] [10, p. 24]. Ausgenommen ist hierbei die gelegentliche Verarbeitung von Daten, die Verarbeitung von nicht sensiblen Daten oder wenn die Verarbeitung von personenbezogenen Daten im strafrechtlichen Kontext geschieht und die Verarbeitung zu keinem Risiko für die Freiheiten und Rechte der natürlichen Personen führt. Außerdem ausgenommen sind Behörden sowie öffentliche Stellen. Der/Die Verantwortliche haftet jedoch weiterhin selbst [32].

In **Artikel 28** werden die Pflichten der Auftragsverarbeiter definiert. Ein Auftragsverarbeiter ist im datenschutzrechtlichen Kontext entweder eine natürliche oder eine juristische Person, Einrichtung oder andere Stelle, welche personenbezogenen Daten im Auftrag eines anderen Verantwortlichen verarbeitet [10, p. 4]. Zusammengefasst werden die Pflichten und Rechte des Auftragsverarbeitenden in schriftlicher Form auf Grundlage eines Vertrages oder eines anderen Rechtsinstruments festgehalten und definiert. Der/Die Auftragsverarbeiter*in darf hierbei ohne schriftliche Genehmigung des/der Verantwortlichen keine weiteren Schritte eigenständig übernehmen. Der/Die Verantwortliche hat sicherzustellen, dass der/die ausgewählte Auftragsverarbeiter*in ausreichende Garantien bietet, dass adäquate technische und organisatorische Maßnahmen vollzogen werden, um im Einklang mit den Anforderungen der DSGVO [10] zu stehen und den Schutz der Freiheiten und Rechte der betroffenen Person zu gewährleisten. Der Auftragsgeber verpflichtet sich durch diesen Vertrag zur Vertraulichkeit und unterliegt einer gesetzlichen Verschwiegenheitspflicht. Nachdem der Vertrag erfüllt wurde oder der/die Verantwortliche den Auftrag kündigt, kann der/die Verantwortliche entscheiden, ob die personenbezogenen Daten zurückgegeben oder gelöscht werden müssen. Die Warnpflicht verpflichtet den/die Auftragsverarbeiter*in im Fall einer Datenschutzverletzung den/die Verantwortliche unverzüglich darüber zu informieren. Damit die Verarbeitung der Daten überprüft

werden kann, muss der/die Auftragsverarbeiter*in alle Informationen zum Nachweis der Einhaltung seiner Pflichten zur Verfügung stellen [10, pp. 24-25] [34].

Artikel 30 verpflichtet die/den Verantwortliche*n dazu, ein Verzeichnis über die Verarbeitung von Daten zu führen. Das Verzeichnis muss alle Verarbeitungstätigkeiten, die in der Zuständigkeit der verantwortlichen Person liegen, führen. Dabei müssen zum Beispiel die Kontaktdaten des Vertreters oder Datenschutzbeauftragen vorhanden sein. Außerdem sind die Zwecke der Verarbeitung, sowie die allgemeine Beschreibung von technischen Schutzmaßnahmen, die durch den Verantwortlichen getroffen wurden, festzuhalten. Das Verzeichnis kann schriftliche oder elektronisch geführt werden und müssen auf Anfrage der Aufsichtsbehörde zur Überprüfung vorgelegt werden [10, pp. 27-28] [35].

Die Sicherheit der Datenverarbeitung von personenbezogenen Daten wird in Abschnitt2, **Artikel 32** geregelt. Dabei wird auf den Stand der Technik geachtet und gleichermaßen die Implementierungskosten bewertet. Damit nach Bewertung des Risikos und der Eintrittswahrscheinlichkeiten ein angemessenes Schutzniveau bei der Verarbeitung der personenbezogenen Daten gewährleistet wird, hat der Verantwortliche laut DSGVO [10] folgende Maßnahmen zu treffen: [10, p. 28] [33]

- Die Pseudonymisierung und Verschlüsselung der personenbezogenen Daten [10, p. 28] [33].
- Die Vertraulichkeit, Integrität und Verfügbarkeit der Systeme müssen im Zusammenhang mit der Datenverarbeitung auf Dauer gewährleistet werden. Dieser Zustand kann beispielsweise durch Zugangskontrollen und Zugriffsbeschränkungen erzielt werden [10, p. 28] [33].
- Bei einem Ausfall muss die Verfügbarkeit und der Zugang zu den personenbezogenen Daten rasch wiederhergestellt werden können. Durch Backup-Software oder Wiederherstellungspunkten kann hier rasch reagiert werden [10, p. 28] [33].
- Ein Verfahren, welches die regelmäßige Überprüfung und Bewertung der Wirksamkeit von technischen und organisatorischen Maßnahmen zur Sicherstellung der Sicherheit der Datenverarbeitung gewährleistet. Das kann durch Selbstevaluierungsprozesse, welche im Datenverarbeitungsprozess eingegliedert sind, erreicht werden [10, p. 28] [33].

Infolge der Beurteilung von Risiken und des angemessenen Schutzniveaus, sind besonders Risiken zu bewerten, die mit der Datenverarbeitung im Zusammenhang stehen. Insbesondere unbeabsichtigte oder unrechtmäßige Veränderungen, Verluste oder Vernichtungen von personenbezogenen Daten sind zu beachten [10, p. 29].

Im Falle einer **Datenschutzverletzung**, welche zu einem Risiko für die Rechte und Freiheiten natürlicher Personen führen können, hat der/die Verantwortliche auf schnellstem Wege und möglichst binnen 72 Stunden, nachdem die Datenschutzverletzung bekannt geworden ist, eine Meldung an die Aufsichtsbehörde zu machen. Außerdem haben die betroffenen Personen informiert zu werden, wenn die Verletzung des Schutzes personenbezogener Daten ein hohes Risiko für die Freiheiten und Rechte der betroffenen Personen darstellt [10, p. 29] [36].

Der sogenannte ***data breach*** kann durch den Verlust eines Datenträgers oder eines Hackerangriffs stattfinden. Hierbei verliert der/die Verantwortliche die Kontrolle über die erhobenen personenbezogenen Daten. Für die betroffenen Personen kann das zu Identitätsdiebstahl, Rufschädigung oder andere wirtschaftliche sowie soziale Nachteile führen [36]. Bei Verletzung der Meldepflicht kann der/die Verantwortliche zu einer Geldstrafe von bis zu 10 Millionen Euro oder im Falle eines Unternehmers zu einer Strafe von 2% des gesamten weltweit erzielten Jahresumsatzes des vorigen Jahres verklagt werden [36]. Bei besonders schweren Vergehen steigt die Strafe auf 20 Millionen Euro oder 4% des gesamten weltweit erzielten Jahresumsatzes.

Artikel 35 definiert die Voraussetzungen und Analyseschritte für die Risikofolgenabschätzung. In der DSGVO [10] werden die Verantwortlichen, anstelle von Vorabkontrollen oder Meldungen an Behörden, zu einer eigenverantwortlichen Evaluierung der Risiken in Bezug auf die geplante Verarbeitung von Daten verpflichtet. Durch die Datenschutz-Folgeabschätzung sollen die Risiken und Auswirkungen auf die Rechte und Freiheiten der Betroffenen analysiert und in der Folge risikomindernde Maßnahmen ergriffen werden [37]. In der DSGVO [10, pp. 30-31] wird festgehalten, dass insbesondere, wenn neue Technologien zur Verarbeitung von personenbezogenen Daten zum Einsatz kommen oder die Art und Zwecke der Verarbeitung möglicherweise ein hohes Risiko für die Rechte und Freiheiten der Betroffenen darstellen, der/die Verantwortliche eine Datenschutzfolgeabschätzung durchzuführen hat. Folgende Fälle sind laut Artikel 35 [10, p. 31] als konkrete Beispiele angeführt:

- Die systematische und umfangreiche Auswertung persönlicher Daten, betreffend wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Verhalten, Aufenthaltsort oder Arbeitsort von betroffenen Personen. Dabei wird insbesondere das *Profiling*, welches als Entscheidungsgrundlage dient und in rechtswirksamen Maßnahmen gegen oder für betroffene Personen enden kann, angesprochen [10, p. 31].
- Die umfangreiche Verarbeitung von sensiblen Daten wie z.B. Fingerabdrücke oder Gesundheitsdaten oder von personenbezogenen Daten in Verbindung mit strafrechtlichen Urteilen [10, p. 31].
- Die Überwachung von öffentlichen Bereichen zum Beispiel durch Videoüberwachung [10, p. 31].

Unter einer systematischen Datenverarbeitung wird die Verarbeitung durch ein organisiertes und methodisch arbeitendes System verstanden. Der/Die Verantwortliche hat in solchen Fällen, bei der Abwicklung der Datenschutz-Folgeabschätzung den Rat des/der nominierten Datenschutzbeauftragten einzuholen. Dabei ist zu beachten, dass die Datenschutzbehörde ein „Black List“ für Fälle, bei denen auf jeden Fall eine Datenschutz-Folgeabschätzung durchgeführt werden muss, erstellt hat. Als Gegenform zur „Black List“ gibt es eine „White List“ für Fälle, in denen keine Folgeabschätzung erstellt werden muss. Die ehemalige Artikel-29-Datenschutzgruppe hat unter anderem neun Kriterien erstellt, die als Entscheidungsgrundlage für datenverarbeitende Stellen gelten. Wobei davon ausgegangen wird, dass bei der Erfüllung von zwei dieser neun Kriterien ein hohes Risiko für die Betroffenen besteht und eine Datenschutz-Folgeabschätzung durchgeführt werden muss [10, pp. 31-32] [37].

In einer **Datenschutz-Folgeabschätzung** [10, pp. 18-19] [25]
sind die Verarbeitungsvorgänge sowie Verarbeitungszecke zu beschreiben. Außerdem ist die Notwendigkeit und die Verhältnismäßigkeit der Verarbeitung zu bewerten. Dabei muss die Rechtmäßigkeit und das berechtige Interesse an der Verarbeitung der Daten der Betroffenen beschrieben werden. Unter anderem müssen in Bezug auf die Betroffenenrechte die Risiken für die Rechte und Freiheiten der Betroffenen erläutert werden. Damit diesen Risiken entgegengewirkt werden kann, müssen Abhilfemaßnahmen und risikomindernde Aktionen gesetzt und definiert werden. Falls der/die Verantwortliche keine Pläne zur Minderung des Risikos für die Betroffenen vorlegen kann und ein hohes Risiko für die Rechte und Freiheiten dieser Personen besteht, hat der/die Verantwortliche die Aufsichtsbehörde zu kontaktieren. Diese hat 8 Wochen Zeit, schriftliche Empfehlungen zur Risikominderung zuzustellen. Damit diese schriftlichen Empfehlungen auf die individuellen Bedürfnisse eines Falles reagieren können, hat der die Verantwortliche Informationen wie z.B. die Zwecke und Mittel der Datenverarbeitung der zuständigen Datenschutzbehörde zu übermitteln.

Artikel 37 verpflichtet die/den Verantwortlichen eine/n Datenschutzbeauftrage*n zu benennen, wenn

- Die Datenverarbeitung durch eine Behörde oder öffentliche Stelle durchgeführt wird [10, pp. 31-32] [37].

- Die Kerntätigkeit in der Verarbeitung von Daten, die eine umfangreichen und regelmäßigen [10, pp. 31-32] [37] Überwachung von betroffenen Personen erforderlich machen, liegt [10, pp. 31-32] [37].
- Die Verarbeitung eine umfangreiche Erfassung von Daten besonderer Kategorien beinhaltet [10, pp. 31-32] [37].
- Die Verarbeitung personenbezogene Daten über strafrechtliche Urteile und Straftaten enthält [10, pp. 31-32] [37].

Der/Die Datenschutzbeauftragte wird aufgrund seiner Qualifikation und seines Fachwissens auf dem Gebiet des Datenschutzrechts ausgewählt. Dabei kann die ausgewählte Person als Beschäftigter des Verantwortlichen agieren oder einen externen Dienstleistungsvertrag erfüllen. Die Kontaktdaten des Datenschutzbeauftragten müssen im Anschluss der Einstellung der Aufsichtsbehörde mitgeteilt werden. Gerichte müssen, im Falle, dass sie in ihrer justiziellen Tätigkeit agieren, keine/n Datenschutzbeauftragte*n einstellen [10, p. 33]. Der/Die Datenschutzbeauftragte hat folgende Aufgaben zu erfüllen: [10, p. 33]

- Die verantwortlichen Personen über ihre Pflichten nach der DSGVO [10] oder anderen Datenschutzrichtlinien zu informieren [10, p. 33].
- Die Umsetzung und Einhaltung der DSGVO [10] sowie anderer Datenschutzrichtlinien überwachen [10, p. 33].
- Die Sensibilität für das Datenschutzthema bei den Mitarbeitern durch Schulungen steigern [10, p. 33].
- Beratung auf Anfrage im Zusammenhang mit der Datenschutz-Folgeabschätzung [10, p. 33].
- Fungiert als Anlaufstelle für externe Behörden wie zum Beispiel die Aufsichtsbehörde [10, p. 33].

In Abschnitt 5, **Artikel 40-43** [10, pp. 38-39] wird definiert auf welche Weise Verbände oder andere Vereinigungen von Verantwortlichen Verhaltensregeln in Bezug auf die faire und transparente Verarbeitung der personenbezogenen Daten festlegen können. Die Kommission trägt die Verantwortung, dass nur den genehmigten Verhaltensregeln allgemeine Gültigkeit zuteilwird und sie in einer geeigneten Form veröffentlicht werden. Die Überwachung der Einhaltung dieser genehmigten Verhaltensregeln wird durch eine Stelle durchgeführt, welche das notwendige Fachwissen besitzt und von der Aufsichtsbehörde zu diesem Zweck akkreditiert wurde. Damit es während der Überwachung zu keinen Interessenskonflikten kommt, ist besonders auf die Unabhängigkeit der überwachenden Stelle zu achten. Die Aufsichtsbehörden fördern laut DSGVO Artikel 42 [10, p. 38] die Einführung von datenschutzspezifischen Gütesiegeln und Zertifizierungen. Mithilfe dessen können die Verantwortlichen die rechtmäßige und korrekte Datenverarbeitung nachweisen. Dabei ist zu beachten, dass die Zertifizierung durch ein transparentes und freiwilliges Verfahren erreicht werden kann. Die Verpflichtungen und Rechte der Zertifizierungsstellen werden in Artikel 43 festgelegt. Dabei spielt zum Beispiel die Unabhängigkeit der Zertifizierungsstellen eine wichtige Rolle.

3.2.5 Wann sind Daten ausreichend anonymisiert?

Damit personenbezogene Daten laut DSGVO [10] ausreichend anonymisiert sind und dadurch nicht mehr in den Rechtsbereich der DSGVO [10] fallen muss die **faktische Anonymisierung** erreicht werden. Diese ist erreicht, wenn die Re-Identifizierung eines Subjekts in einem Anonymitätssets nur durch völlig unverhältnismäßige Maßnahmen möglich wäre [38, p. 4]. Dabei verhindert sie nicht in jedem Fall eine Re-Identifizierung. Laut DSGVO [10] ist das Schutzniveau der faktischen Anonymisierung jedoch ausreichend.

Dabei verpflichtet sich der/die Verantwortliche und der/die Auftragsverarbeiter*in unter Berücksichtigung des aktuellen Stands der Technik und den damit verbundenen Kosten dazu, angemessene organisatorische und technische Maßnahmen wie zum Beispiel die Pseudonymisierung zu treffen, um die faktische Anonymisierung zu gewährleisten.

Damit sind die Daten ausreichend anonymisiert, dass sie für statistische Auswertungen oder Forschungsarbeiten verwendet werden können und nicht mehr in den Rechtsbereich der DSGVO [10] fallen. Der Angreifer müsste in dem Fall der faktischen Anonymisierung einen derartigen hohen Kosten-, Zeit- und Arbeitsaufwand zur Re-Identifizierung betreiben, dass der Aufwand den möglichen Nutzen der erbeuteten Daten überwiegt [39, pp. 7-8].

3.3 USA

In den Vereinigten Staaten gibt es bis heute, kein umfassendes Gesetz zur Regelung der Datenerhebung, des Datenschutzes und der Privatsphäre. Es gibt eine Anzahl von Bundes- und Landesgesetzen, die für bestimmten Sektoren gelten und dort die Verarbeitung von persönlichen Informationen regeln [40].

In den USA entstanden im Jahr 1789 die ersten Ansätze des Datenschutzes durch den **vierten Zusatzartikel** der Verfassung der Vereinigten Staaten. Dieser schränkte die Befugnisse der Regierung, zur willkürlichen Durchsuchung von Häusern und der Beschlagnahmung von Gegenständen der Menschen in den USA, ein. Beamte der Regierung mussten eine richterliche Genehmigung einholen, um Häuser durchsuchen zu dürfen [41, p. 5] [42].

Im Jahr 1890 entstand ein Artikel zur Überprüfung eines Gesetzes Namens **Right to Privacy (oder the right to be let alone)** [41, p. 10] [42], welches von Richter Louis Brandeis als auch Samuel Warren geschrieben wurde. Es ist einer der einflussreichsten Aufsätze der amerikanischen Rechtsgeschichte und gilt als die erste Veröffentlichung in den USA, in der ein Recht auf Privatsphäre gefordert wird. Zu dieser Zeit waren Sofortbildaufnahmen und Zeitungsunternehmen in den privaten und häuslichen Bereich des Lebens eingedrungen. Die beiden Autoren prüften, ob das bestehende Recht einen Grundsatz bietet, auf den man sich zum Schutz der Privatsphäre des Einzelnen berufen kann.

1914 wurde durch den **FTCA (Federal Trade Commission Act)** [40] [42] die **FTC (Federal Trade Comission)** gegründet, die betrügerischen Geschäftspraktiken verbot. Dabei ist die FTC seit den 1970er Jahren die führende Bundesbehörde, welche sich am häufigsten mit Fragen des Datenschutzes, der Datensicherheit, den dazugehörigen Regulatorien und ihrer Durchsetzung befasst. Die rechtlichen Befugnisse zur Durchsetzung erhält die FTC aus Abschnitt 5 des FTCAs, der irreführenden Praktiken auf dem wirtschaftlichen Markt verbietet und die Unternehmen zwingt sich an Datenschutzrichtlinien zu halten. Die Bundesbehörde ist ebenfalls in der Lage zivilrechtliche Geldstrafen durch eine Reihe von Gesetzen durchzusetzen. Die FTC bietet Maßnahmen zu einer breiten Anzahl an Datenschutzproblemen an. Darunter fallen Spam, Probleme in sozialen Netzwerken, verhaltensorientierte Werbung, Mobilfunkbetrug.

Im Jahr 1960 wurden durch **William L. Prosser** vier Datenschutzdelikte veröffentlicht, welche die heutige Gesetzeslage zum Datenschutz in den USA maßgeblich beeinflusst haben. Falls die Privatsphäre einer Person durch eines dieser Delikte gestört werden sollte, räumte Prossers Artikel dem/der Geschädigten das Recht ein, den/die Täter/in auf Schadensersatz zu verklagen. Dabei handelt es sich bei der ersten rechtswidrigen Handlung, um den Schutz vor vorsätzlichem physischem oder anderweitem Eindringen in die Privatsphäre des Opfers. Der zweite Punkt beschäftigt sich mit der Veröffentlichung von privaten Details. Falls die veröffentlichte Angelegenheit für eine Person beleidigend oder das veröffentlichte Material nicht von öffentlichem Interesse ist, ist die Veröffentlichung unzulässig. Der dritte Punkt sagt aus, dass es unrechtmäßig ist, eine andere Person in ein falsches Licht zu stellen und frei erfundene beleidigende Fakten über das Opfer zu veröffentlichen. Der letzte Punkt deklariert Identitätsdiebstahl als strafbare Handlung [41, pp. 14-16] [42].

1967 bezeichnete **Alan Westin** die Privatsphäre als den Anspruch einer Einzelperson selbst zu bestimmen, wann, wie und in welchem Umfang Informationen über sie weitergegeben darf. Sein Buch trug maßgeblich dazu bei, die Rahmbedingungen für eine moderne Debatte über Technologie, Privatsphäre und persönliche

Freiheit zu setzen [42]. Zwischen 1960 und 1970 gab es bahnbrechende Entscheidungen vor Gericht, welche die Privatsphäre der Bürger weiter bestärkte. Ein Beispiel ist „Katz v. United States (1976)“ in dem ein Grundsatzurteil des Obersten Gerichtshofs den Schutz des vierten Zusatzartikels ausweitete. Dadurch waren Bürger vor unrechtmäßigen Durchsuchungen, Abhörungen und Beschlagnahmungen, über die Wohnung hinaus, auf alle Orte, an denen sie eine angemessene Erwartung auf Privatsphäre hatten, geschützt [41, p. 22] [42].

Das **FERPA (Family Educational Rights and Privacy Act)** von 1974, auch genannt der „Buckley Amendment“, ist ein Bundesgesetz, dass die Bildungsdatensätze von SchülerInnen und StudentInnen schützt. Dabei geht es um die rechtmäßige Einhaltung der Datenschutzanforderungen in Bezug auf den Schutz von personenbezogenen Daten und Verzeichnisinformationen. Es berechtigt die SchülerInnen und ihre Eltern auf ihre Datensätze zugreifen zu können, Änderungen an diesen anzufordern oder die Offenlegung der Informationen steuern zu können. Das Gesetz gilt für alle US-Bildungseinrichtungen die Bundesmittel vom US-Bildungsministerium erhalten [40] [41, p. 27] [42].

Das am 31.Dezember 1974 in Kraft getretene Datenschutzgesetz **Privacy Act** [41, pp. 26-27], ist ein US-Bundesgesetz, dass die Richtlinie für eine zulässige Sammlung, Verarbeitung und Verbreitung von personenbezogenen Daten durch Bundesbehörden festlegt. Er gibt Einzelpersonen das Recht, ihre Daten einzusehen und diese, wenn notwendig, korrigieren zu lassen. Der Privacy Act gilt jedoch nicht für den privaten Sektor, noch für lokale oder staatliche Behörden. Lediglich Bundesbehörden sind verpflichtet die Richtlinien des Privacy Acts einzuhalten. Außerdem sind personenbezogene Daten, welche durch routinemäßige Verwendung offengelegt werden, solange sie zu dem Zweck offengelegt werden, zu dem sie gesammelt wurden, vom Schutz des Privacy Acts ausgenommen. Das bildet ein großes Schlupfloch für die Verarbeitung und Weitergabe von personenbezogenen Daten durch Bundesbehörden, da der Begriff „routinemäßige Verwendung“ dehnbar ist. Der Privacy Act zielte darauf ab, die Verwendung der SSN (Social Security Number) als universelle Identifikationsnummer durch Behörden zu verringern. Dadurch dass der Privacy Act die Verwendung der SSN als Identifikation von Personen im privaten Sektor nicht regelte, setzte sich der steigende Trend durch. Die SSN ist eine neunstellige Nummer, welche zur Identifikation an alle US-Bürger und berechtigte Personen durch die US-Regierung vergeben wird. Sie wird bei der Eröffnung eines Bankkontos oder beim Kauf eines Autos benötigt. Außerdem kann durch die SSN die gesamte Arbeitszeit oder das Einkommen einer Person eingesehen werden.

Der **TCPA (Telephone Consumer Protection Act)** [40] [41, p. 36] [42], auch bekannt als der „Kennedy-Kassebaum Act“, wurde im Jahr 1986 in Kraft gesetzt. Er ermöglichte es den Bürgern, sich in ein „Do Not Call Registry“ eintragen zu lassen. Daraufhin durften Telemarketingunternehmen diese Personen nicht anrufen. Ansonsten konnten die Personen die Unternehmen auf bis zu 500 Dollar pro Anruf verklagen. Damit sollten automatische Telefonanrufe und bestimmte Arten von Werbeanrufen unterbunden werden und den Verbrauchern die Möglichkeit gegeben werden sich davor zu schützen. Obwohl es sich hierbei nicht um ein Online-Datenschutzgesetz handelt, ist es eines der bekanntesten Datenschutzgesetzte in den Vereinigten Staaten.

Der **HIPAA (Health Insurance Portability and Accountability Act)** von 1996 wurde in Kraft gesetzt, um den Informationsfluss im Gesundheitswesen zu regulieren. Das Ziel des HIPAA ist es, die Verarbeitung und Weitergabe von Gesundheitsdaten im Gesundheitswesen und der Versicherungsbranche vor Diebstahl und Missbrauch zu schützen. In die Kategorie der Gesundheitsdaten fallen alle Gesundheitsinformationen, welche mit einer Person in Verbindung gebracht werden können. Das Gesetz schreibt außerdem vor, dass jegliche Weitergabe der Gesundheitsdaten, welche nichts mit der Behandlung, der Bezahlung oder dem Betrieb des Gesundheitswesens zu tun hat, durch die betroffene Person genehmigt werden muss. Strafverfolgungsbehörden erhalten Zugang zu Gesundheitsdaten, wenn sie diese zum Zwecke der Identifizierung oder der Auffindungen eines Verdächtigen oder einer vermissten Person benötigen [40] [41, pp. 37-38] [42].

Der COPPA (Childrens Online Privacy Protection Act) [40] [41, p. 38] [42] wurde am 21.Oktober 1998 als Bundesgesetz in den USA im Kongress verabschiedet und trat im April 2000 in Kraft. Das Gesetz regelt die Erfassung von personenbezogenen Daten von Kindern unter 13 Jahren im Internet. Es verpflichtet Betreiber von Websites oder Online-Diensten, welche sich an Kinder unter 13 Jahren richten oder Kenntnis davon haben, dass sie personenbezogene Daten von Kindern sammeln, bestimmte Maßnahmen zum Schutz der Daten zu treffen. Darunter fällt, dass in den Datenschutzrichtlinien geregelt sein muss, wie die Zustimmung der Erziehungsberechtigten zur Erhebung, Verwendung und Weitergabe der personenbezogenen Daten ihrer Kinder überprüft wird. Unteranderem sind die Erziehungsberechtigten ermächtigt die Löschung der personenbezogenen Daten ihrer Kinder zu beantragen.

Der 1999 verabschiedete **GLBA (Gramm-Leach-Bliley Act)** [41, p. 39] [42] ist ein Bundesgesetz, dass die Weitergabe von nicht öffentlichen personenbezogenen Kundendaten durch Finanzinstitute an ihre verschiedenen Tochtergesellschaften oder Zweigstellen, welche unterschiedliche Dienstleistungen erbringen, reguliert. Die Unternehmen, welche an dem Informationenaustausch beteiligt sind, müssen ihre Kunden über die Weitergabe der Daten informieren. Das Gesetz verpflichtet die Finanzinstitute, die Weitergabe von Kundendaten zu erläutern, den Kunden die Möglichkeit zu geben, die Weitergabe ihrer Daten abzulehnen und die privaten Daten durch einen vom Institut erstellten Sicherheitsplan zu schützen. Die Umsetzung des Gesetzes wurde von der „Federal Trade Commission's Privacy of Consumer Financial Information Rule“ (Privacy Rule), den Bundesbanken und anderen Bundesaufsichtsbehörden sowie den staatlichen Versicherungsaufsichtsbehörden vorangetrieben und durchgesetzt. Es führte zu einer Massenversendung von Datenschutzerklärungen an die Kunden, in denen sie über ihre Rechte informiert wurden.

Das **E-Government-Gesetz** [42] aus dem Jahr 2002 wurde vom Kongress in dem Bestreben verabschiedet, die Bundesverwaltung ins 21. Jahrhundert zu bringen, indem Fortschritte in der Informationstechnologie genutzt werden, um den Zugang und die Nutzung von Behördendiensten zu verbessern. Ein zentraler Bestandteil des Gesetzes war die Vorschrift, dass alle Bundesbehörden eine Datenschutz-Folgeabschätzung (Privacy Impact Assessment, PIA) für jede neue Technologie durchführen müssen, die personenbezogene Daten sammelt, verwaltet oder verbreitet.

Im Jahr 2003 war **Kalifornien** der erste Staat, der Gesetze zur Meldung von Datenschutzverletzungen einführte. Die neue Gesetzgebung verpflichtete Unternehmen und staatliche Behörden zur Meldung, wenn persönliche Daten von Bürgern aus Kalifornien bei einer Datenpanne preisgegeben wurden. Die meisten anderen US-Bundesstaaten und einige Länder im Ausland haben ihre Gesetze zur Offenlegung von Datenschutzverletzungen an diese Gesetzgebung angelehnt [42]. Bis Anfang 2006, hatte fast die Hälfte der Bundesstaaten ähnliche Gesetze zur Offenlegung von Datenschutzverletzungen erlassen. Im Jahr 2008 schufen die FTC und die NCUA (National Credit Union Administration) die „Red Flags Rule“. Die Regel sollte dazu beitragen Identitätsdiebstahl zu verhindern. Obwohl sie bereits im Januar 2008 verabschiedet wurde, verzögerte sich ihre Durchsetzung aufgrund des Widerstands der Opposition bis zum 31. Dezember 2010 [42].

Ein weiteres nationales Gesetz ist das kalifornische Gesetz zum Schutz der Privatsphäre von Verbrauchern. **CCPA (California Consumer Privacy Act)** ist ein staatliches Gesetz, das regeln soll, wie Unternehmen mit den personenbezogenen Daten der Einwohner des Bundesstaates Kalifornien umgehen. Das CCPA wurde 2018 verabschiedet und trat am 1. Januar 2020 in Kraft. In den darauffolgenden Jahren unterzeichneten die Gouverneure der Bundesstaaten Virginia und Colorado ebenfalls Datenschutzgesetze zum Schutz ihrer Bürger. In den folgenden Jahren werden immer mehr Bundesstaaten, getrieben durch die rasante Entwicklung von neuen Technologien, Datenschutzgesetze zum Schutz ihrer Bürger zu erlassen [41, p. 46] [42].

4 Anonymisierung

4.1 Einführung

In der heutigen Zeit ist das Bedürfnis nach aussagekräftigen Daten stärker denn je. Private sowie öffentliche Einrichtungen aus fast allen Branchen, benötigen für die Weiterentwicklung und Forschung in ihrem Bereich immer mehr Daten. Dabei ist ein Konkurrenzkampf um die Vorherrschaft der Daten, welche wie in Kapitel 3 beschrieben, rechtlichen Regulatorien unterliegen, entfacht. Es ist anzumerken, dass in einigen dieser Datenverarbeitungsvorhaben, nicht personenbezogene Daten zur Erfüllung der Ziele ausreichend Informationen liefern. Falls die Identifikation der Person, von der die Daten verarbeitet werden, nicht notwendig ist, so eignen sich anonymisierte Daten als Datenbasis für umfassende statistische Analysen oder zu Forschungszwecken [39, p. 2].

Der Vorteil, der Nutzung von anonymen Daten in der EU ist, dass diese Daten nicht mehr in den Rechtsbereich der DSGVO [10] fallen und somit frei genutzt werden können. Durch aktuelle Anonymisierungstechniken kann bereits eine absolute Anonymisierung von personenbezogenen Daten erzielt werden. In solchen Fällen ist es möglich eine Re-Identifikation von einzelnen Individuen zur Gänze auszuschließen. Nachdem die Anonymisierung eines Datensatzes mit personenbezogenen Daten abgeschlossen ist, darf trotz Veröffentlichung der Daten, keine Re-Identifizierung von einzelnen Personen mehr möglich sein [16, pp. 27-28].

4.2 Anonymität & Pseudonymität

Damit das Themengebiet der Anonymisierung samt der dazugehörigen Begriffe verständlich ist, folgt eine Auflistung mehrerer Begriffe und ihrer Bedeutung.

Die **Identität** [16, p. 12] eines Objekts oder einer Person besteht aus einer großen Menge an Attributen, durch diese die Identität eines Objekts oder einer Person in einer großen Gruppe von Objekten oder Personen identifizierbar ist. Die Gesamtanzahl der Personen einer beobachteten Menge kann unterschiedlich sein. Zum Beispiel kann es, wie in dieser Arbeit durchgeführt, das Ziel sein, eine einzelne Taxi-ID aus einer Gruppe von mehreren Taxi-IDs durch einen individuellen Attributwert zu identifizieren. Dabei kann ein Subjekt verschiedene Identitäten besitzen. Ein Student, der in einem großen Universitätssaal einer großen Menge an Studenten beiwohnt, kann in diesem Fall einer Menge, die einer statistischen Auswertung betreffend Studenten als Datenbasis dient, zugerechnet werden. Auf der anderen Seite kann derselbe Student einer Menge von Bewohnern eines Studentenheims, die einer Forschungsarbeit als Datenbasis dient, zugerechnet werden. Dasselbe Subjekt wird in der ersten statistischen Auswertung als Student gesehen und in der zweiten Forschungsarbeit als Bewohner. Dabei ist das einzelne Subjekt eindeutig identifizierbar, wenn es sich in einer Menge von Subjekten ausreichend von den anderen Subjekten unterscheidet. Weiters ist zwischen der vollständigen und der partiellen Identität zu unterscheiden. Die vollständige Identität beinhaltet alle Attribute eines Subjektes. Auf die vollständige Identität haben im Normalfall nur sehr wenige Personen Zugriff. Die partielle Identität spiegelt nur einen Teil der vollständigen Identität wider. Dabei enthält sie nur einen gewissen Teil der Gesamtmenge an Attributen. Im vorhin genannten Beispiel ist es möglicherweise so, dass das Subjekt in der ersten statistischen Auswertung die Attribute eines Studenten inne hat. Bei der zweiten Forschungsarbeit ist es möglich, dass das Subjekt ausschließlich die Attribute eines Bewohners besitzt. Wenn nun die Attribute des Subjekts ganzheitlich betrachtet werden, sind die beiden Teilmengen ein Bruchteil von der Gesamtmenge der Attribute des Subjekts [16, p. 12].

Wenn ein einzelnes Subjekt in einer Menge von Subjekten hingegen nicht ausreichend identifizierbar ist, so wird das als **Anonymität** bezeichnet. In diesem Kontext bedeutet nicht ausreichend identifizierbar, dass der

Angreifer das Subjekt in einer Menge von Subjekten, auch genannt Anonymitätsset, nicht ausreichend von anderen Subjekten unterscheiden kann [6, p. 8]. Als Angreifer kann eine Person oder eine Stelle bezeichnet werden, die es sich als Ziel gesetzt hat, einzelne Subjekte aus einem Anonymitätsset zu identifizieren. In dieser Arbeit wird eben dieser Vorgang dokumentiert und das Ergebnis festgehalten. Dabei spielt die Gesamtanzahl von Subjekten eines Anonymitätssets bezogen auf das Anonymitätslevels, eine entscheidende Rolle. Wenn von zwei Anonymitätssets, s1 und s2, ausgegangen wird. Das erste Set(s1) besteht aus drei Subjekten. Im Vergleich dazu besteht das zweite Set(s2) aus zehn Subjekten. Hierbei kann davon ausgegangen werden, dass die Subjekte in Set zwei(s2), falls beide Sets mithilfe desselben Anonymitätsverfahrens anonymisiert wurden, ein höheres Level an Anonymität bietet [6, p. 8]. Der Vorgang der Anonymisierung bezeichnet eine Reihe von Verfahren, um personenbezogene Daten so weit zu verändern, dass die Attribute keiner einzelnen Person in dem Anonymitätsset mehr zugeordnet werden können und dadurch nicht mehr in den Rechtsbereich der DSGVO [10] fallen. Bei der Wahl eines geeigneten Anonymitätsverfahrens ist zu beachten, dass erstens die verarbeiteten Daten nur bis zum erforderlichen Anonymitätsgrad anonymisiert werden und zweitens das Analysepotential der Daten für statistische Auswertungen nicht verloren geht. Für realitätsnahe Auswertungen sind realitätsnahe Daten notwendig [16, pp. 12-13].

Die **Pseudonymisierung** wird zur Verschleierung der wahren Identität einer Person verwendet. Hierbei wird der Name einer Person durch ein Pseudonym z.B. eine Kennung ersetzt. Ein Pseudonym kann zum Beispiel ein Nickname in einem Online-Forum sein. Die *Pseudonymität* wird erreicht, wenn der Angreifer durch das Pseudonym nicht mehr auf die reale Identität des Subjekts schließen kann. Die Pseudonymisierung beschreibt die Veränderung der personenbezogenen Daten mittels Referenztabelle. Dadurch kann der Angreifer ohne Wissen über die Referenztabelle keinen direkten Personenbezug herstellen. In der Referenztabelle wird jedem Subjekt das dazugehörige Pseudonym zugewiesen. Deswegen ist es von hoher Wichtigkeit, dass diese Referenztabelle nicht in die falschen Hände gerät [16, pp. 13-14].

Dabei lässt sich die Anonymisierung in zwei unterschiedlich starke Anonymisierungslevel aufteilen. Bei der **absoluten Anonymisierung** von personenbezogenen Daten, ist es unabhängig davon, ob der Angreifer Zusatzwissen von Dritten besitzt, unmöglich durch derzeit verfügbare Methoden eine Re-Identifizierung von Subjekten durchzuführen. Die absolute Anonymisierung ist laut Erwägungsgrund 26 der DSGVO [10] nicht nötig. Dabei ist anzumerken, dass nachdem die Daten einen Zustand der absoluten Anonymität erreicht haben, diese in den meisten Fällen für statistische Forschungszwecken unbrauchbar sind.

Die **faktische Anonymisierung** ist erreicht, wenn die Re-Identifizierung eines Subjekts in einem Anonymitätssets nur durch völlig unverhältnismäßige Maßnahmen möglich wäre [38, p. 4]. Dabei gewährleistet sie keine absolute Anonymität und verhindert nicht in jedem Fall eine Re-Identifizierung. Laut DSGVO [10] ist die faktische Anonymisierung jedoch ausreichend. Damit sind die Daten ausreichend anonymisiert, dass sie, wie in Abbildung 3 zu sehen, für statistische Auswertungen oder Forschungsarbeiten verwendet werden können. Der Angreifer müsste in dem Fall der faktischen Anonymisierung einen derartigen hohen Kosten-, Zeit- und Arbeitsaufwand zur Re-Identifizierung betreiben, dass der Aufwand den möglichen Nutzen der erbeuteten Daten überwiegt [39, pp. 7-8].

Die Artikel-29-Datenschutzgruppe hat 2014 in einer Stellungnahme, einen Maßstab für eine ausreichend faktische Anonymisierung von Daten bereitgestellt. Da die verschiedenen Anonymisierungsverfahren ein unterschiedliches Schutzniveau gegenüber Re-Identifizierung bieten, hat die Datenschutzgruppe drei Risiken bzw. Angriffsmethoden vorgestellt, gegen welche ein Anonymisierungsverfahren schützen sollte, damit die Daten als faktisch anonym gelten [39, p. 8] [43, p. 13].

1. *Singling out (Herausgreifen)* – ist die Option, ein oder mehrere Subjekte in einem Datenbestand auszumachen [39, p. 8] [43, p. 13].

2. *Verknüpfbarkeit* – Wenn der Angreifer in der Lage ist, zwei Datensätze des gesuchten Subjekts in unterschiedlichen Datenbestände miteinander zu verknüpfen, ist von Verknüpfbarkeit die Rede. Sobald der Angreifer durch eine Korrelationsanalyse die Verknüpfbarkeit zwischen zwei unabhängigen Datenbeständen hergestellt hat aber dabei kein einzelnes Subjekt identifizieren kann, bietet das Anonymisierungsverfahren Schutz vor *Herausgreifen*, jedoch nichtmehr vor der *Verknüpfbarkeit* [39, p. 8] [43, p. 13].
3. *Inferenz* – Wenn der Angreifer aus dem vorliegenden Gesamtdatenbestand Informationen über das gesuchte Subjekt ableiten kann, spricht man von Inferenz. Die personenbezogene Inferenz tritt auf, wenn der Angreifer mit einer bestimmten Wahrscheinlichkeit den Wert eines personenbezogenen Merkmals, durch das Ableiten von Merkmalen anderer Datensätze des Datenbestands, ableiten kann [39, p. 8] [43, p. 13].

Durch den Schutz gegen die drei genannten Risiken kann mit einem gewissen Grad an Sicherheit die faktische Anonymität der Daten garantiert werden. Damit ist die Privatsphäre der Subjekte des Datenbestands gewährleistet und die datenverarbeitende Stelle kann die Daten für weitere Zwecke nutzen. Falls die drei genannten Risiken nicht vollständig ausgeschlossen werden können, muss der/die Datenverarbeiter*in eine Evaluierung bezüglich des Risikos der Identifizierung eines Subjekts durchführen. Nach der Evaluierung ist durch den/die Datenverarbeiter*in das Restrisiko zu berücksichtigen und Gegenmaßnahmen vorzunehmen [39, pp. 8-9].

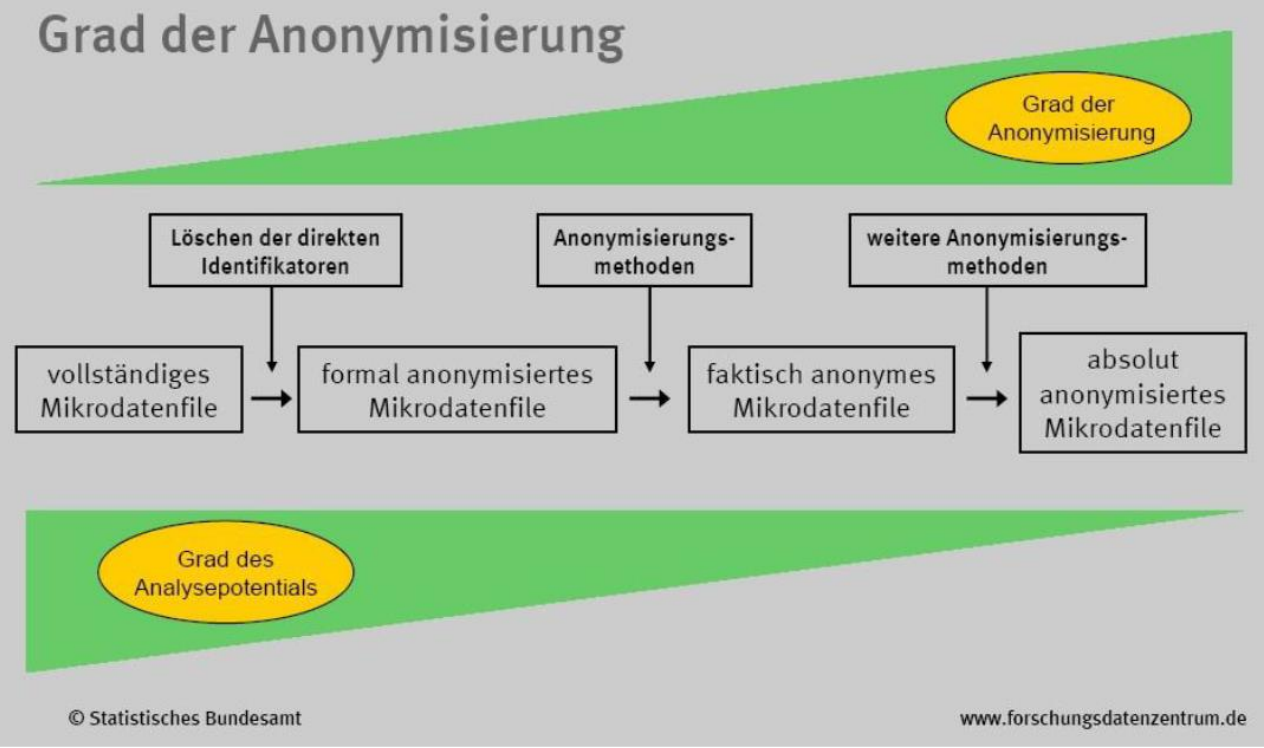

Abbildung 3: Grad der Anonymisierung [44]

Wie in Abbildung 3 zu sehen ist nimmt mit einem höheren Grad der Anonymisierung das Analysepotential schrittweise ab. Nach der **formalen Anonymisierung,** in der lediglich direkte Identifikationsmerkmale, wie zum Beispiel Name, Anschrift oder Telefonnummer einer Person verändert werden, ist der Grad des Analysepotentials ausreichend hoch. Je mehr die Daten jedoch verändert und gelöscht werden, desto niedriger ist das Analysepotential [44].
Für die Anonymisierung bestehen zwei Überbegriffe:

1. **Datenverändernde Methoden** – zufällige Veränderung der Datenbasis.
2. **Datenaggregierende Methoden** – die Aggregation von Daten.

4.3 Mikrodaten

Ein Mikrodatensatz besteht aus mehreren Datensätzen die eine einzelne Person, ein Unternehmen oder anderweitige Subjekte darstellen. Diese Daten werden zum Beispiel aus Umfragen, Studien oder Volkszählungen gewonnen. Dabei müssen die einzelnen Zusammenhänge zwischen den Merkmalen bedacht und evaluiert werden. Ein Beispiel hierzu wäre der Mikrodatensatz von Personen und deren Krankheitsbilder. Nach der Anonymisierung der Daten kann durch nachlässige Betrachtung der Zusammenhänge von Krankheitsbildern und Geschlecht, durch manche Krankheiten ein Rückschluss auf das Geschlecht der erkrankten Person getroffen werden.

Eine Tabelle aus Mikrodaten kann wie in Abbildung 8 zu sehen ist aus verschiedenen Arten von Merkmalen bestehen. Einige dieser Attribute führen zur direkten Identifizierung der Subjekte im Mikrodatensatz. Andere hingegen können durch Kombination mit Zusatzwissen zur Identifikation des Subjekts beitragen [45, pp. 13-14].

4.4 Arten von Variablen

Bei **qualitativen Variablen** auch genannt **kategoriale Variablen** handelt es sich um Variablen, welche eine endliche Anzahl an Werten oder Kategorien annehmen können. Dabei beschreiben sie die Zugehörigkeit eines Subjekts zu einer definierten Kategorie [45, pp. 4-5]. Bei exakt zwei Möglichkeiten spricht man von binären qualitativen Variablen, wie zum Beispiel die Ausprägung Schwanger und Nicht-Schwanger. Im Normalfall sind qualitative Variablen beschreibender Natur und keine Zahlen. Weitere Beispiele sind Geschlecht, Wohnort, Haarfarbe [46].

Bei **quantitativen Variablen** auch genannt **numerische Variablen** handelt es sich um Variablen, welche eine unendliche Anzahl an natürlichen und reellen Zahlen annehmen können. Der Vorteil von quantitativen Variablen ist die Möglichkeit, dass sie als Datenbasis für arithmetische Operationen verwendet werden können. Beispiele, die durch quantitative Variablen beschrieben werden, sind Gewicht, Gehalt, Alter, Schulnoten [46].

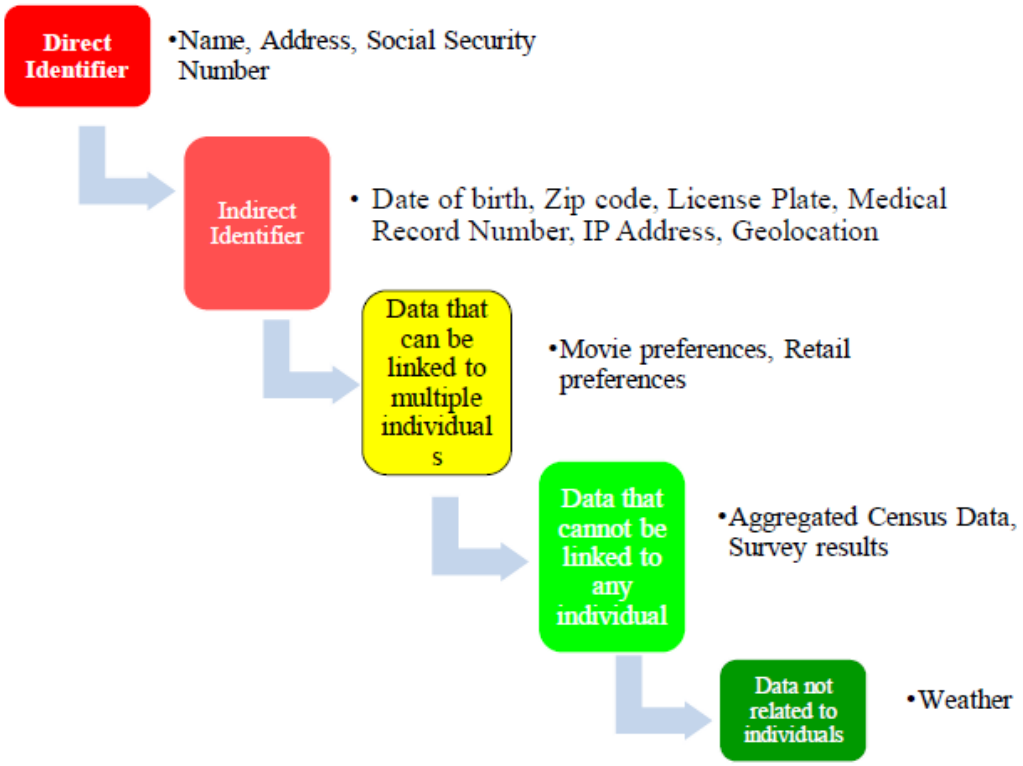

Abbildung 4: Identifikatoren [47, p. 204]

Eine weitere Klasse der Variablen nennt sich **Identifikatoren**. Diese lassen sich in einem stufenartigen System, wie in Abbildung 7, je nach Aussagekraft darstellen. An oberste Stelle befinden sich Variablen wie die Sozialversicherungsnummer, Handynummer oder ein Name. Diese Variablen tragen zur direkten Identifikation einer Person bei und sollten daher durch das Anonymisierungsverfahren entfernt werden. Im Fachjargon wird diese Klasse von Identifikatoren als **direkte Identifikatoren** bezeichnet [47, p. 203] [48, p. 106].

An zweiter Stelle befinden sich die **indirekten Identifikatoren (Quasi-Identifikatoren).** Mithilfe dieser Art von Identifikatoren können Personen zwar nur indirekt, aber trotzdem eindeutig identifiziert werden. Laut [16, p. 30] [47, p. 203] muss ein Angreifer, um eine Person in einem Datensatz identifizieren zu können, entweder eine geringe Anzahl an indirekten Identifikatoren kennen oder durch Kombination von Zusatzwissen und indirekten Identifikatoren den Bezug herstellen. Beispiele von indirekten Identifikatoren sind zum Beispiel Geburtsdatum, IP-Adresse, Geschlecht, Wohnbezirk [48, p. 106].
Wie in [49, p. 22] angeführt, können 87% (216 Millionen von 248 Millionen) Menschen in den USA durch die Kombination der Quasi-Identifikatoren Geschlecht, Geburtsdatum und 5-stelligen ZIP Code identifiziert werden. Weiters wird in [49, p. 22] angeführt, dass 53% (132 Millionen von 248 Millionen) Menschen durch die Kombination der Quasi-Identifikatoren Wohnort, Geschlecht und Geburtsdatum identifiziert werden. Für diese Auswertungen wurden Informationen von Volkszählungen aus den USA im Jahr 1990 herangezogen [49, p. 22].

Die dritte Stufe beschäftigt sich mit Merkmalen, die mehreren Menschen zugeordnet werden können. Darunter fallen zum Beispiel die Körpergröße, Lieblingsfilme oder Lieblingsrestaurant. Die vierte und fünfte Stufe betrifft Daten, die entweder den meisten oder keinem Menschen zugeordnet werden können, wie zum Beispiel das Ergebnis einer Onlineumfrage oder das Wetter [47, p. 203].

Eine kritische Art von Variablen sind **sensitive Attribute.** Dabei handelt es sich um vertrauliche Attribute, die bei Veröffentlichung und Zuordnung, der betroffenen Person in mehreren Hinsichten schaden können. Dabei kann es zu Rufschädigung oder dem Eindringen in die Privatsphäre der betroffenen Person kommen. Beispiele wären hier ein Krankheitsbefund, die Sexualität, das Gehalt oder die politische Gesinnung [48, p. 106].

Als abschließendes Beispiel sind in Abbildung 8 die in den vorigen Absätzen erklärten Arten von Attributen dargestellt. In Anlehnung an [50, p. 4] ist Abbildung 8 als Mikrodatensatz mit dem vollständigen Namen als Direkten Identifikator, dem Geschlecht, Geburtsdatum und der Wohnpostleitzahl als Quasi-Identifikatoren sowie der politischen Gesinnung als sensitivem Attribut zu verstehen. Da der vollständige Name ein direkter Identifikator ist, kann dadurch jede Person im Datensatz eindeutig identifiziert werden. Bei den Quasi-Identifikatoren Geschlecht, Geburtsdatum oder PLZ kann durch gewisse Kombination auf eine einzelne Person rückgeschlossen werden. Das sensitive Attribut, die politische Gesinnung ist in dieser Datenbank das Wertvollste von allen. Durch Anonymitätsverfahren müssten im ersten Schritt die direkten Identifikatoren entfernt werden, um den sofortigen Rückschluss auf eine Einzelperson zu verhindern. Die Quasi-Identifikatoren müssen im weiteren Schritt durch verschiedene Techniken wie dem Aggregieren oder dem Vertauschen so weit verschleiert werden, dass die politische Gesinnung der einzelnen Personen nicht mehr identifizierbar ist.

Direkter Identifikator	Quasi-Identifikator			Sensibles Attribut
Vollständiger Name	Geschlecht	Geburtsdatum	PLZ	Politische Gesinnung
Josef Meier	M	01.01.1980	1010	SPÖ
Marie Holzer	W	06.01.1987	1010	SPÖ
Sarah Thompson	W	15.07.1990	1010	ÖVP
Tobias Rend	M	19.08.1991	1050	ÖVP
Ruth Talin	W	03.09.1970	1050	SPÖ
Sebastian Eisberger	M	24.03.1994	1060	NEOS
Jasmin Scheuer	W	30.12.1990	1060	GRÜNE

Abbildung 5: Mikrodatensatz [50, p. 4]

Als Beispiel kann der Datenschutzskandal der österreichischen Post AG erwähnt werden. Die Post hat seit 2001 einen eigenen Online-Shop zum Handel mit Daten eingerichtet. Neben Name, Adresse, Geschlecht und Alter werden noch weitere Daten gesammelt und den Kunden zu Marketingzwecken verkauft. Des Weiteren hat die Post AG durch statistische Hochrechnungen 2,2 Millionen Österreichern eine mutmaßliche politische Orientierung zugeordnet. Diese Daten wurden unter anderem an österreichischen Parteien verkauft, damit diese zielgerichtet Wahlwerbung versenden können. Die politische Gesinnung fällt jedoch, wie im vorherigen Absatz erwähnt unter den Schutz der DSGVO Artikel 9 [10]. Der Artikel 9 wird in dieser Arbeit im Kapitel 3.2.2 behandelt und erläutert [51].

4.5 Offenlegung (Disclosure)

Bei der Offenlegung von sensitiven Attributen geht es dem Angreifer um den Wissenszugewinn. Dabei ist es das Ziel des Angreifers sein Wissen über das gesuchte Objekt zu erweitern. Das könnte ein Student sein, der durch Bewegungsdaten auf die Taxi-ID der gesuchten Fahrt rückschließen will. Es könnte aber auch ein staatlicher Akteur sein, der zur Einschüchterung durch das Analysieren von Mikrodaten sensitive Informationen über einen kritischen Journalisten sammeln möchte. Bei der Offenlegung von sensitiven Attributen wird zwischen drei Arten unterschieden [6, p. 10] [45, pp. 39-40].

4.5.1 Attribut Offenlegung (Attribute Disclosure)

Bei der *Attribute Disclosure* kann der Angreifer auch ohne direkter Zuordnung einer Person zu einem bestimmten Datensatz sein Wissen über ein sensibles Attribut der Person in dem Datensatz erweitern. Dabei ist es ausreichend, wenn der Angreifer seine Vermutungen durch die Offenlegung des sensitiven Attributs bestärken kann [48, p. 106]. Dabei reicht es aus, wenn der Angreifer bestimmen kann, ob sich der gesuchte Datensatz in der Datenbank befindet [6, p. 10]. Durch dieses Wissen kann der Angreifer weitere Schlüsse ziehen. Ein Beispiel wäre hier eine Datenbank mit SPÖ-Mitgliedern. Allein durch das Wissen, dass sich die gesuchte Person in der Datenbank befindet, weiß der Angreifer, dass sie der SPÖ angehört.

Direkter Identifikator	Quasi-Identifikator			Sensibles Attribut
Vollständiger Name	Geschlecht	Geburtsdatum	PLZ	Politische Gesinnung
Josef Meier	M	01.01.1980	1010	SPÖ
Marie Holzer	W	06.01.1987	1010	SPÖ
Sarah Thompson	W	15.07.1990	1050	ÖVP
Tobias Rend	M	19.08.1997	1050	ÖVP

Abbildung 6: Attribute Disclosure [50, p. 4]

Abbildung 9 soll einen Mikrodatensatz mit den Quasi-Identifikatoren Geburtsdatum und Postleitzahl, sowie dem sensitiven Attribut politische Gesinnung darstellen. Die Attribute Name und Geschlecht werden gelöscht, damit keine direkte Identifikation stattfinden kann. Wenn der Angreifer nun aber bereits weiß, dass seine gesuchte Person im fünften Bezirk wohnt, kann er sofort rückschließen, dass die Person der ÖVP gesinnt ist. Beide Datensätze der Personen aus dem fünften Bezirk haben als politische Gesinnung die ÖVP eingetragen und sind somit anfällig für die Offenlegung der politischen Gesinnung. Diesen Vorgang nennt man die Attributs Offenlegung (Attribute Disclosure) [6, p. 11].

4.5.2 Identität Offenlegung (Identity Disclosure)

Bei der Offenlegung der Identität kann der Angreifer die gesuchte Person exakt einem Datensatz im anonymisierten Datensatz zuordnen [48, p. 106]. Dabei spielt es keine Rolle, ob der Angreifer neues sensibles Wissen über die gesuchte Person erlangt. Allein die Möglichkeit der Re-Identifizierung einer einzelnen Person im Datensatz zeigt auf, dass das gewählte Anonymisierungsverfahren nicht ausreicht, um die Sicherheit und Anonymität der Personen im Datensatz zu gewährleisten [6, p. 10]. Ein Beispiel von *Identity Disclosure* ist die Offenlegung des ehemaligen Gouverneurs aus Massachusetts. Dabei konnte durch die Kombination des Wählerverzeichnisses mit den angeblich anonymisierten und veröffentlichten Gesundheitsdaten durch die Group Insurance Commission (GIC) die Krankheitsdaten des ehemaligen Gouverneurs offengelegt werden. Durch die spezielle Kombination aus Geschlecht, Geburtsdatum und 5-stelligen ZIP-Code wurde der Gouverneur in den anonymisierten Daten identifiziert [49, p. 52].

Direkter Identifikator	Quasi-Identifikator			Sensibles Attribut
Vollständiger Name	Geschlecht	Geburtsdatum	PLZ	Politische Gesinnung
Josef Meier	M	01.01.1980	1010	SPÖ
Marie Holzer	W	06.01.1987	1010	SPÖ
Sarah Thompson	W	15.07.1990	1010	ÖVP
Tobias Rend	M	19.08.1991	1050	ÖVP
Ruth Talin	W	03.09.1970	1050	SPÖ
Sebastian Eisberger	M	24.03.1994	1050	NEOS
Jasmin Scheuer	W	30.12.1954	1070	GRÜNE

Abbildung 7: Identity Disclosure [50, p. 4]

Abbildung 7 ist als Mikrodatensatz mit dem vollständigen Namen als direkten Identifikator, dem Geschlecht, Geburtsdatum und der Wohnpostleitzahl als Quasi-Identifikatoren sowie der politischen Gesinnung als sensitivem Attribut zu verstehen. Da der vollständige Name eine direkter Identifikator ist, kann dadurch jede Person im Datensatz eindeutig identifiziert werden. Bei den Quasi-Identifikatoren Geschlecht, Geburtsdatum oder PLZ kann durch gewisse Kombination auf eine einzelne Person rückgeschlossen werden. Angenommen der Angreifer weiß, dass die gesuchte Person im siebten Bezirk wohnhaft ist, so könnte er in diesem Datensatz sofort auf die letzte Reihe des Datensatzes rückschließen und die gesuchte Person identifizieren. Dieser Vorgang nennt sich *Identity Disclosure.* In diesem Beispiel würde im Zuge der Offenlegung der Identität gleichzeitig eine Offenlegung der politischen Gesinnung erfolgen. Die Kombination darauf belegt, dass es in solchen Fällen im Zuge der *Identity Disclosure* auch zur *Attribute Disclosure* kommt [6, p. 11].

4.5.3 Mitgliedschaft Offenlegung (Membership Disclosure)

Bei der Offenlegung der Zugehörigkeit einer Person zu einer bestimmten Gruppe oder Gemeinschaft spricht man von der *Membership Disclosure.* Dabei kann es für eine Person zum Nachteil kommen, wenn offengelegt wird, dass sie mit einer hohen Wahrscheinlichkeit im Datensatz vorkommt. Ein Beispiel ist ein systemkritischer Journalist. Angenommen dieser schreibt in einer Diktatur, welche ihre Kritiker verfolgen und einsperren lässt, systemkritische Zeitungsartikel für eine freie Zeitung. Da er diese unter einem Pseudonym verfasst und in den anderen Bereichen auch auf seine Anonymität achtet, währt er sich in Sicherheit. Wenn das Zeitungsunternehmen eine Personaldatenbank führt und diese dann durch einen Hackerangriff gestohlen und veröffentlicht wird, so muss der bis jetzt anonym gebliebene Journalist um sein Leben fürchten. In diesem Zusammenhang reicht allein die Offenlegung seiner Mitgliedschaft bei der systemkritischen Zeitung, um sein Leben in Gefahr zu bringen [50, p. 14].

4.6 Datenverändernde Methoden (Perturbative Methods)

Durch datenverändernde Methoden werden die Merkmale einzelner Datensätze eines Datenbestandes mithilfe eines vorher festgelegten Musters verändert. Dabei werden die Daten dahingehend verfälscht, dass es für den Angreifer nicht mehr möglich ist, eine direkte Beziehung zwischen den Daten und des dazugehörigen Subjekts herzustellen. Je nachdem wie die Randomisierung durchgeführt wird, erzeugt der Vorgang ein unterschiedliches Maß an Anonymität. Eine Möglichkeit wäre, die bestehenden Werte einer Kategorie durch ein Wahrscheinlichkeitsmodell mit anderen möglichen Merkmalswerten auszutauschen. Dabei ist es das Ziel, dass die Zusammenhänge der Datensätze weitestgehend erhalten bleiben. Das zufällige Multiplizieren oder Addieren von Werten sind zwei weitere Möglichkeiten zur Randomisierung [39, pp. 10-11] [43, p. 14].

Name	Geschlecht	Alter	Taxi-ID	Gehalt
Josef	M	20	1001	1500€
Marie	W	21	1002	2000€
Sarah	W	22	1003	2100€
Tobias	M	23	1004	1500€
Ruth	W	24	1005	2300€
Sebastian	M	26	1006	2600€
Jasmin	W	28	1007	1700€
Johann	M	29	1008	2300€
Fabienne	W	30	1009	1800€

Abbildung 8: Beispiels Datenbank

Abbildung 8 soll eine einfache Datenbank mit Name, Geschlecht und Gehalt darstellen. Dabei werden in den folgenden Absätzen die verschiedenen Arten der Randomisierung durchgeführt und erklärt.

4.6.1 PRAM (Post-Randomization Method)

Die erste Möglichkeit geht von einem Austausch des Name-Merkmals aus. Angenommen bei männlichen Datensätzen liegt die Wahrscheinlichkeit bei 30 Prozent, dass der Name auf Alexander umgeändert wird. Bei einer Wahrscheinlichkeit von 20 Prozent wird der Name auf Christian geändert und bei einer Wahrscheinlichkeit von 40 Prozent auf Julian. Dahingegen entsteht eine Restwahrscheinlichkeit von 10 Prozent, dass der Name eines männlichen Datensatzes nach dem Vorgang der Randomisierung denselben Wert wie vor der Randomisierung besitzt. Diesen Vorgang nennt man die Post-Randomization-Method [52, pp. 7-8] [52, p. 10].

4.6.2 Addieren & Multiplizieren (Adding Noise)

Bei der zufälligen Addition von Werten können die nummerischen Originalwerte verschleiert werden. In der Abbildung 8 zum Beispiel, kann so zu jedem Gehalts Merkmal ein gewisser Betrag addiert werden, damit das echte Gehalt nicht mehr rekonstruierbar ist. Wenn davon ausgegangen wird, dass zu jedem Gehalts Merkmal ein Betrag von 250€ hinzugefügt wird, so ist nach der Randomisierung nicht mehr rekonstruierbar, wie viel jede Person verdient. Es können lediglich Schätzungen oder Vermutungen abgegeben werden [39, pp. 10-11] [43, p. 14] [52, p. 10].

Name	Geschlecht	Alter	TaxiID	Randomisiertes Gehalt
Josef	M	21	1001	1750€
Marie	W	22	1002	2250€
Sarah	W	23	1003	2350€
Tobias	M	24	1004	1750€
Ruth	W	25	1005	2550€
Sebastian	M	26	1006	2850€
Jasmin	W	27	1007	1950€
Johann	M	28	1008	2550€
Fabienne	W	29	1009	1800€

Abbildung 9: Randomisierte Datenbank (*Adding Noise*)

Wie in der randomisierten Datenbank in Abbildung 9 zu sehen ist, sind die Gehälter um den zufällig gewählten Betrag erhöht worden. Der Vorteil dieser Art der Randomisierung ist, dass durch das gleichmäßige Erhöhen der Gehälter die Zusammenhänge zwischen den Daten erhalten bleiben. Nach wie vor besitzt der Taxilenker namens Sebastian das höchste Gehalt. Der Mitarbeiter namens Josef hingegen hat trotz der randomisierten Gehaltserhöhung das niedrigste Gehalt. Der Angreifer kann also das genaue Gehalt der einzelnen Personen im Umkehrschluss nicht mehr bestimmen. Er kann nur zu einer gewissen Wahrscheinlichkeit davon ausgehen, dass das Gehalt von Sebastian zwischen 2500 und 3000 Euro liegt. Die dritte Möglichkeit wäre das Multiplizieren der Gehälter mit einem vordefinierten Faktor. Wie auch bei der Addition bleibt der Zusammenhang zwischen den Daten trotz Randomisierung erhalten. Die Personen mit den höchsten und niedrigsten Gehälter sind nach wie vor auszumachen. Randomisierung allein reicht daher nicht aus, um die Personen einer Datenbank vollständig vor Re-Identifizierung zu schützen. In den meisten Fällen erschwert die Randomisierung nur das Re-Identifizieren bis zu einem gewissen Grad. Daher muss sie mit Generalisierungsmethoden verstärkt werden [39, pp. 10-11] [43, p. 14].

4.6.3 Vertauschung (Data Swapping)

Beim *Data Swapping* [45, pp. 32-33] oder auch *Vertauschung* genannt, werden die Werte eines Merkmals verschiedener Datensätze in einem Datensatz miteinander vertauscht. Das hat den Vorteil, dass die Verteilung der Werte in der Gesamtmenge erhalten bleiben, jedoch die ursprünglichen Zusammenhänge

zwischen Merkmalswerten und Subjekten nicht mehr rekonstruierbar sind. Außerdem ist dieser Vorgang der Anonymisierung auch auf qualitative Merkmale anwendbar. Der Nachteil besteht darin, dass ein Angreifer, da die Werte auch nach der *Vertauschung* ihren Originalwert entsprechen, exakte Informationen über, wie in diesem Fall, Gehaltsdaten erhält. Falls nun der Angreifer die Person mit dem höchsten Gehalt in dem Taxiunternehmen kennt, kann er durch personenbezogene Inferenz herausfinden, wie hoch das exakte Gehalt der Person ist.

Name	Geschlecht	Alter	TaxiID	Randomisiertes Gehalt
Josef	M	20	1001	2100€
Marie	W	21	1002	2000€
Sarah	W	22	1003	1500€
Tobias	M	23	1004	2300€
Ruth	W	24	1005	1500€
Sebastian	M	26	1006	1700€
Jasmin	W	28	1007	2600€
Johann	M	29	1008	2300€
Fabienne	W	30	1009	1800€

Abbildung 10: Randomisierte Datenbank (*Data Swapping*) [45, pp. 32-33]

Wie in Abbildung 10 zu sehen ist, sind die einzelnen Gehälter der TaxifahrerInnen zufällig miteinander vertauscht worden. Dabei ist die Verteilung der Gehälter gleichgeblieben und es ist nach wie vor möglich, eine statistische Auswertung zum Durchschnittsgehalt im Taxiunternehmen durchzuführen. Das Ergebnis wird, so wie auch vor dem Vorgang des *Data Swappings,* das Gleiche sein. Der Nachteil, der im Absatz davor beschrieben wurde, ist nun gut nachvollziehbar. Wenn der Angreifer das Zusatzwissen besitzt, dass der Taxilenker Sebastian das höchste Gehalt der Taxifahrer im Unternehmen besitzt, so weiß er trotz *Data Swapping,* dass Sebastian exakt 2600 Euro verdient.

4.6.4 Mikroaggregation (Microaggregation)

Bei der Mikroaggregation werden mehrere kleine Mikrocluster von ähnlichen Datensätzen in einer Datenbank gebildet. Dabei werden im ersten Schritt die betroffenen Subjekte zu einer Gruppe zusammengefasst und im zweiten Schritt werden die Merkmale durch das arithmetische Mittel der Gruppe ersetzt. Die Clustergruppe sollte dabei mindestens aus 3 einzelnen Datensätzen bestehen. Bei einem Cluster von nur 2 Datensätzen ist die Wahrscheinlichkeit der Re-Identifikation zu hoch [6, pp. 200-201] [52, pp. 8-9].

Name	Geschlecht	Alter	TaxiID	Gehalt
Josef	M	20	1001	1625€
Fabienne	W	30	1009	1625€
Jasmin	W	28	1007	1625€
Tobias	M	23	1004	1625€
Ruth	W	24	1005	2260€
Sebastian	M	26	1006	2260€
Johann	M	29	1008	2260€
Marie	W	21	1002	2260€
Sarah	W	22	1003	2260€

Abbildung 11: Mikroaggregation (*Microaggregation*) [52, pp. 8-9]

Wie in Abbildung 11 zu sehen ist wurden die einzelnen TaxifahrerInnen in Gehaltscluster aufgeteilt. Dabei wurde für jedes Cluster das Durchschnittsgehalt berechnet und als einheitliches Gehalt angegeben. Im

Vergleich zu Abbildung 8 kann nach der Mikroaggregation kein Rückschluss auf das exakte Gehalt der einzelnen Personen getroffen werden. Die Personen wurden in zwei Klassen eingeteilt. Klasse 1 sind alle Personen mit einem Gehalt weniger als 2000€. Klasse 2 hingegen enthält alle Personen die mehr als 2000€ verdienen. Danach wurde das Durchschnittsgehalt je Klasse berechnet und die Mikrocluster gebildet [6, pp. 200-201].

4.6.5 Runden (Rounding)

Beim Runden werden quantitative Merkmale verändert. Durch den Vorgang sollen die Werte auf ein Vielfaches des Ursprungswertes multipliziert und dann zum Beispiel auf die nächstgelegenen 10er Stelle gerundet werden. Im Experiment dieser Diplomarbeit wurden die Start- und Endkoordinaten der Taxifahrten bis zur 3. Stelle gerundet. Somit ist die Anzahl der Taxi-IDs pro Gruppe, die dieselben Start- und Endkoordinaten haben, gestiegen [45, p. 30].

Name	Geschlecht	Alter	TaxiID	Gehalt
Josef	M	20	1001	3000€
Marie	W	21	1002	4000€
Sarah	W	22	1003	4000€
Tobias	M	23	1004	3000€
Ruth	W	24	1005	5000€
Sebastian	M	26	1006	5000€
Jasmin	W	28	1007	3000€
Johann	M	29	1008	5000€
Fabienne	W	30	1009	4000€

Abbildung 12: Runden (*Rounding*) [45, p. 30]

In Abbildung 12 wurden die Gehälter der einzelnen Datensätze mit 2 multipliziert und danach auf die nächste tausender Stelle gerundet. Somit ist das Originalgehalt verschleiert und der Angreifer kann lediglich Annahmen über das Originalgehalt treffen.

4.7 Datenaggregierende Methoden (Non-Perturbative Methods)

4.7.1 Generalisierung (Generalisation)

Die Generalisierung von Daten wird hauptsächlich auf Quasi-Identifikatoren angewendet, dabei werden einzelne Subjekte eines Datensatzes zu größeren Gruppen (Äquivalenzklassen) zusammengefasst. Qualitative Attribute werden bei der Generalisierung verallgemeinert und durch einen Überbegriff getauscht. Quantitative Attribute können durch die Darstellung eines Intervalls zusammengefasst werden. Dabei ist es nach der Generalisierung des Alter Merkmals für den Angreifer nicht mehr möglich, das exakte Alter einer einzelnen Person in dem Datensatz zu bestimmen. Trotz alledem kann der Angreifer die Gruppe, in der sich die gesuchte Person nach der Generalisierung befindet, beobachten und mit einer gewissen Wahrscheinlichkeit das gesuchte Merkmal bestimmen [6, p. 219].

Bei der Generalisierung wird zwischen *Global recoding* und *Local recoding* unterschieden. Durch *Global recoding* werden alle Werte eines Attributs auf den gleichen Detailgrad generalisiert. In Abbildung 13 würden daher alle Ingenieure und Anwälte als Akademiker, sowie alle Sänger und Maler als Künstler gelten. Da das Merkmal Beruf bei allen Datensätzen einer Datenbank durch *Global recoding* generalisiert wird, ist der Informationsverlust am größten [6, p. 219].

Beim *Local recoding* werden die Merkmale eines jeden Datensatzes einzeln oder für jede Äquivalenzklasse extra generalisiert. Dabei kann zum Beispiel für die Personen, die zwischen 30-35 Jahre alt sind, entschieden

werden, dass das Merkmal Beruf entweder auf Akademiker oder Künstler generalisiert wird. Auf der anderen Seite kann für Personen, zwischen 35 und 40 entschieden werden, dass sie weiterhin als Sänger, Maler, Ingenieure oder Anwälte bezeichnet werden. Dadurch kann je nach Bedarf der Grad der Anonymisierung bestimmt und der Informationsverlust so niedrig wie möglich gehalten werden. *Top and Bottom Coding* wird ein Generalisierungsmechanismus bezeichnet, in welchem Ausreißer-Werte anonymisiert werden. Das kann zum Beispiel in einer Personaldatenbank das Merkmal Gehalt betreffen. Hierbei werden alle Personen mit einem exorbitanten Gehalt generalisiert. Genauso müssen Personen, die im Vergleich zum Durschnitt ein sehr geringes Gehalt verdienen verschleiert werden. Dieser Mechanismus verstärkt die Anonymisierung und schützt gegen Re-Identifizierung [6, pp. 219-220].

Wenn sich durch den Anonymisierungsvorgang 5 Personen in der Altersklasse von 20-25 befinden, kann der Angreifer mit einer 20-prozentigen Wahrscheinlichkeit das gesuchte Gehalt bestimmen. Nach der Generalisierung der Daten sind alle individuellen Subjekte einer Gruppe zugeordnet und die K-Anonymität erreicht. Der Buchstabe K steht in diesem Fall für die Menge der kleinsten Subjektgruppe der Datenbank. In Kapitel 4.4.1 wird auf die K-Anonymität noch genauer eingegangen und ein Beispiel zum besseren Verständnis dargelegt [39, pp. 11-12] [43, p. 19].

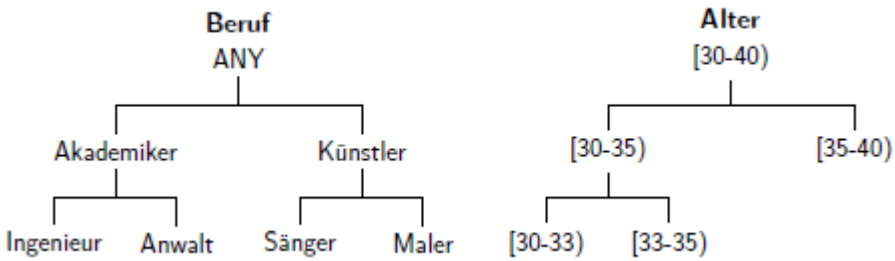

Abbildung 13: Generalisierung von Merkmalen [50, p. 25]

In Abbildung 13 wird erkenntlich, dass durch die Generalisierung des Merkmals Beruf ein Ingenieur oder Anwalt zum Überbegriff Akademiker zusammengefasst werden kann. Kreative Berufe wie der eines Sängers oder Malers fallen in den Überbegriff Künstler. Das Alter ist wie in Abbildung 13 dargestellt als quantitatives Merkmal auf einer Intervallskala abzubilden. Dadurch gehen zwar exakte Informationen der einzelnen Datensätze einer Datenbank verloren, jedoch erhöht der Vorgang der Generalisierung die Privatsphäre der einzelnen Personen [50, p. 25].

4.7.2 Unterdrückung (Suppression)

Zusätzlich zur Generalisierung wird die *Suppression* [50, p. 26] [6, p. 220] verwendet. Dabei werden die zu unterdrückenden Werte als unterdrückt markiert. Die *Suppression* findet hauptsächlich bei qualitativen Merkmalen statt. Dabei gibt es unterschiedliche Ausprägung wie zum Beispiel die *Record Suppression.* Dabei wird ein ganzer Dateneintrag unterdrückt und als solcher markiert. Bei der *Value Suppression* werden hingegen nur einzelne Werte unterdrückt. Außerdem gibt es bei der Unterdrückung genauso wie bei der Generalisierung ein *Local Suppression* und eine *Global Suppression.* Der Unterschied wurde in Kapitel 4.7.1 bei der Generalisierung erläutert.

Name	Geschlecht	Alter	TaxiID	Randomisiertes Gehalt
Josef	M	21	1001	1750€
Marie	W	22	1002	2250€
Sarah	W	23	1003	2350€
Tobias	M	24	1004	1750€
Ruth	W	25	1005	(missing)
Sebastian	M	26	1006	(missing)
Jasmin	W	27	1007	1950€
Johann	M	28	1008	(missing)
Fabienne	W	29	1009	1800€

Abbildung 14: Unterdrückung (*Suppression*) [6, p. 220]

In Abbildung 14 werden die drei höchsten Gehälter durch die *Suppression* unterdrückt und der in Kapitel 4.7.1 erläuterte Ansatz des *Top Codings* verwendet. Dabei werden die höchsten Gehälter unterdrückt, damit die ausreißenden Werte nicht zur Re-Identifizierung beitragen.

4.8 Anonymitätskriterien

Damit der Grad der Anonymität messbar gemacht werden kann, gibt es verschiedene Methoden. Wie die k-Anonymität funktioniert und die dazugehörigen Erweiterungen werden in den nachfolgenden Kapiteln erläutert.

4.8.1 k-Anonymität (k-Anonymity)

Den Grundbaustein zur Beschreibung der Anonymität legten Sweeney und Samarati im Jahr 1998 mit der Erfindung der k-Anonymität. Dabei wird die Offenlegung der Identität einer Einzelperson (*Identity Disclosure)* verhindert, indem jede Wertekombination von Quasi-Identifikatoren mindestens k Personen eindeutig zugeordnet wird. Dieser Status wird zum Beispiel durch Generalisierungsmethoden in Kombination mit Unterdrückungsmethoden erreicht [50, p. 15] [53, p. 4] [54, p. 8]. Noch bevor die k-Anonymität erreicht werden kann, müssen im ersten Schritt alle expliziten Identifikatoren entfernt werden. Kapitel 4.4 erläutert diese Art von Attributen genauer. Im nächsten Schritt werden die Quasi-Identifikatoren dahingehend generalisiert und unterdrückt, dass für jede Äquivalenzklasse k idente Einträge übrig bleiben. Als Endprodukt ist es das Ziel der k-Anonymität, die Entfernung des Personenbezugs und den Schutz vor Re-Identifizierung [53, pp. 4-5].

Direkter Identifikator	Quasi-Identifikator			Sensibles Attribut
Vollständiger Name	Geschlecht	Alter	PLZ	Politische Gesinnung
Josef Meier	M	20	1010	SPÖ
Marie Holzer	W	21	1020	SPÖ
Sarah Thompson	W	22	1030	ÖVP
Tobias Rend	M	23	1040	ÖVP
Ruth Talin	W	24	1050	SPÖ
Sebastian Eisberger	M	25	1060	NEOS
Jasmin Scheuer	W	26	1070	GRÜNE
Gernot Flick	M	27	1080	GRÜNE
Friedrich Scheuer	M	28	1090	GRÜNE

Abbildung 15: Originaldatensatz [50, p. 4]

In Abbildung 15 ist der Originaldatensatz zu sehen mit dem direkten Identifikator Name, den Quasi-Identifikatoren Geschlecht, Alter und Postleitzahl sowie dem sensitiven Attribut der politischen Gesinnung. Nach Schritt 1 ist in Abbildung 16 zu sehen, dass der direkte Identifikator entfernt wurde, um die sofortige Re-Identifizierung einer Einzelperson zu verhindern.

Quasi-Identifikator			Sensibles Attribut
Geschlecht	Alter	PLZ	Politische Gesinnung
M	20	1010	SPÖ
W	21	1020	SPÖ
W	22	1030	ÖVP
M	23	1040	ÖVP
W	24	1050	SPÖ
M	25	1060	NEOS
W	26	1070	GRÜNE
M	27	1080	GRÜNE
M	28	1090	GRÜNE

Abbildung 16: Explizite Identifikatoren entfernt

Trotz der Entfernung der direkten Identifikatoren kann durch die Kombination von Quasi-Identifikatoren der Personenbezug wiederhergestellt werden. Wenn der Angreifer weiß, dass sein gesuchte Ziel männlich und jünger als 23 ist, so weiß er, dass seine gesuchte Person ein SPÖ-Anhänger ist. Deshalb werden durch die Generalisierungsmethode Äquivalenzklassen gebildet, um die *k-Anonymität* zu erreichen. In diesem Fall muss zusätzlich zur Bildung der Äquivalenzklassen das Merkmal Geschlecht unterdrückt werden, um den Personenbezug zu verhindern. Die Quasi-Identifikatoren Alter und Postleitzahl werden, da sie beide quantitative Merkmale sind, durch eine Intervallskala zusammengefasst.

Quasi-Identifikator			Sensibles Attribut
Geschlecht	Alter	PLZ	Politische Gesinnung
*	20-22	1010-1030	SPÖ
*	20-22	1010-1030	SPÖ
*	20-22	1010-1030	ÖVP
*	23-25	1040-1060	ÖVP
*	23-25	1040-1060	SPÖ
*	23-25	1040-1060	NEOS
*	26-28	1070-1090	GRÜNE
*	26-28	1070-1090	GRÜNE
*	26-28	1070-1090	GRÜNE

Abbildung 17: Äquivalenzklassen k=3, Unterdrückung Geschlecht

Wie in Abbildung 17 zu sehen ist, gibt es jeweils 3 idente Kombinationen von Quasi-Identifikatoren pro Äquivalenzklasse. Somit ist nicht mehr nachvollziehbar, welche Person welche politische Gesinnung besitzt. Das Ziel der k=3 Anonymität ist damit erreicht und der Datensatz anonymisiert. Dabei ergeben sich bestimmte Schwachstellen, die zum Beispiel in Abbildung 17 in den letzten drei Reihen zu beobachten sind. Obwohl die Personen generalisiert und das Geschlecht unterdrückt wurde, kann der Angreifer das sensible Attribut der politischen Gesinnung für alle Personen, die zwischen 26-28 Jahre alt sind und im 7,8 oder 9 Bezirk wohnen, identifizieren. Der Rückschluss ist ein Ergebnis der fehlenden Diversität, da die Einträge der letzten Äquivalenzklasse alle die gleiche politische Gesinnung besitzen. Deshalb schützt die k-Anonymität nur vor *Identity Disclosure* und nicht vor *Attribute Disclosure.*

4.8.2 l-Diversität (l-Diversity)

Aus den in Kapitel 4.8.1 erwähnten Schwachstellen ergibt sich eine weitere Schutzmethode, die die k-Anonymität unterstützen soll. Die l-Diversität besagt, dass sich pro Äquivalenzklasse l unterschiedliche sensitive Attribute befinden müssen [55]. Das hat zur Folge, dass Schwachstellen wie in Abbildung 17 behoben werden und die Re-Identifizierung durch die l-Diversität weiter erschwert wird. Die Werte müssen laut [55] „gut repräsentiert" sein. Dabei wird die Bedeutung von „gut repräsentiert" bewusst offen gelassen. Es gibt hierzu verschiedene Arten der l-Diversität wie zum Beispiel die *Entropy l-Diversity* oder die *Recursive (c, l) – Diversity.* [55] Diese Methoden definieren mithilfe der Entropie, die Häufigkeit, in welcher das sensible l-Attribut in der Datenbank vorkommen muss.

Direkter Identifikator	Quasi-Identifikator			Sensibles Attribut
Vollständiger Name	Geschlecht	Alter	PLZ	Politische Gesinnung
Josef Meier	M	20	1060	SPÖ
Marie Holzer	W	31	1020	SPÖ
Sarah Thompson	W	22	1070	ÖVP
Tobias Rend	M	33	1030	ÖVP
Ruth Talin	W	44	1150	SPÖ
Sebastian Eisberger	M	45	1160	KPÖ
Jasmin Scheuer	W	26	1080	GRÜNE
Gernot Flick	M	37	1040	GRÜNE
Friedrich Scheuer	M	48	1170	GRÜNE

Abbildung 18: Originaldatensatz [50, p. 4]

In Abbildung 18 ist der Originaldatensatz zu sehen mit dem direkten Identifikator Name, den Quasi-Identifikatoren Geschlecht, Alter und Postleitzahl sowie dem sensitiven Attribut der politischen Gesinnung. Nach Schritt 1 ist in Abbildung 19 zu sehen, dass der direkte Identifikator entfernt wurde, um die sofortige Re-Identifizierung einer Einzelperson zu verhindern.

Quasi-Identifikator			Sensibles Attribut
Geschlecht	Alter	PLZ	Politische Gesinnung
M	20	1060	SPÖ
W	31	1020	SPÖ
W	22	1070	ÖVP
M	33	1030	ÖVP
W	44	1150	SPÖ
M	45	1160	KPÖ
W	26	1080	GRÜNE
M	37	1040	GRÜNE
M	48	1170	GRÜNE

Abbildung 19: Explizite Identifikatoren entfernt

Als nächster Schritt werden wie in Kapitel 4.8.1 durch Generalisierungsmethode Äquivalenzklassen gebildet, um die *k-Anonymität* zu erreichen. In diesem Fall muss zusätzlich zur Bildung der Äquivalenzklassen das Merkmal Geschlecht unterdrückt werden, um den Personenbezug zu verhindern. Die Quasi-Identifikatoren Alter und Postleitzahl werden, da sie beide quantitative Merkmale sind, durch eine Intervallskala zusammengefasst. Zusätzlich wird in diesem Beispiel nun eine 3-Diversität angestrebt. Das bedeutet pro Äquivalenzklasse müssen 3 verschiedene sensible Attribute vorkommen.

Quasi-Identifikator			Sensibles Attribut
Geschlecht	Alter	PLZ	Politische Gesinnung
*	20-26	1060-1080	SPÖ
*	20-26	1060-1080	ÖVP
*	20-26	1060-1080	GRÜNE
*	30-37	1020-1040	ÖVP
*	30-37	1020-1040	SPÖ
*	30-37	1020-1040	GRÜNE
*	44-48	1150-1170	SPÖ
*	44-48	1150-1170	KPÖ
*	44-48	1150-1170	GRÜNE

Abbildung 20: l=3 Diversität, k=3 Anonymität

Wie in Abbildung 20 zu sehen ist, wurde die 3-Diversität und 3-Anonymität erreicht. Pro Äquivalenzklassen gibt es 3 Einträge mit identen Quasi-Identifikatoren sowie 3 unterschiedliche sensible Merkmalswerte. Je größer die l-Diversität in Kombination der k-Anonymität ist, desto höher ist der Anonymitätsgrad. Da die l-Diversität die k-Anonymität impliziert, wird sie mit k = l definiert. Anzumerken ist, dass durch die Bildung größerer Äquivalenzklasse ein Informationsverlust unumgehbar ist. Daher sollte die notwendige Anonymität vor dem Anonymitätsverfahren definiert und festgehalten sowie die Folgen für den Detailgrad der Datenbank evaluiert werden [55]. Eine Schwachstelle der l-Diversität ist, dass verschiedene Werte eines sensiblen Attributs eine ähnliche Aussagekraft über die Identität der Person besitzen können. In der Abbildung 20 ist in der letzten Äquivalenzklasse zwar die l-Diversität erreicht worden aber die politische Gesinnung aller Personen dieser Klasse kann als „eher links“ eingestuft werden [48, pp. 108-109].

4.8.3 t-Nähe (t-Closeness)

Aus den in Kapitel 4.8.2 erwähnten Schwachstellen ergibt sich eine weitere Schutzmethode, die die *l-Diversity* unterstützen soll. Die *t-Closeness* besagt, dass sich die Verteilung der sensiblen Merkmalswerte der *l-Diversity* konformen Äquivalenzklassen von der Gesamtverteilung der Datenbank möglichst wenig unterscheidet. Das Ziel von *t-Closeness* ist den Informationszugewinn des Angreifers so niedrig wie möglich zu halten. Eine Äquivalenz Klasse besitzt *t-Closeness*, wenn der Abstand zwischen der Verteilung eines sensiblen Attributs in dieser Klasse und der Verteilung des Attributs in der gesamten Tabelle nicht größer ist als Schwellenwert t. Eine Tabelle besitzt *t-Closeness*, wenn alle Äquivalenzklassen Klassen *t-Closeness* haben [48, p. 109]. Auf die Berechnungsformel der *t-Closeness* wird in dieser Arbeit nicht weiter eingegangen. In der Arbeit von Ninghui Li et al. [48] werden Berechnungsmethoden vorgestellt, eine davon ist die *Earth Mover's distance (EMD).*

Zum besseren Verständnis wurde in [48] ein Theorie Beispiel durchgespielt. Dabei beginnt alles mit einem Angreifer der eine gewisse Vermutung *(A0)* über ein sensibles Attribut seiner gesuchten Person hat. Im nächsten Schritt bekommt der Angreifer eine generalisierte Datenbank, in der alle Quasi-Identifikatoren zu allgemeineren Definition zusammengefasst wurden. Die Vermutungen des Angreifers werden durch die Veröffentlichung der Datenbank und ihrer sensiblen Werte (*Q)* beeinflusst und verändern sich zu *A1*. Der Angreifer weiß über die Quasi-Identifikatoren seines Ziels Bescheid und kann dadurch die gesuchte Äquivalenzklasse in der Datenbank ausmachen. Deshalb weiß er über die Verteilung *P* der sensiblen Attribute in der Äquivalenzklasse seiner gesuchten Person Bescheid und seine Vermutungen ändern ihren Status zu *A2.*

Die *l-Diversity* versucht den Informationsgewinn des Angreifers zwischen *A0* und *A2* so gering wie möglich zu halten. Das gelingt durch die erforderliche Diversität der sensiblen Attribute jeder Äquivalenzklasse.

Ninghui Li et al. [48] gehen davon aus, dass die Inhalte der Veröffentlichung der sensiblen Attribute *Q* aller Personen der Datenbank, öffentlich zugängliche Informationen sind. Dabei soll das *t-Closeness* Modell den Informationsgewinn über bestimmte Personen im Speziellen mindern. Angenommen durch Generalisierung und Unterdrückung werden alle Quasi-Identifikatoren so weit verallgemeinert, dass diese kaum zur direkten Re-Identifizierung beitragen. Die Veröffentlichung der sensiblen Merkmale *Q* ist der Grund für die Veröffentlichung. Das bedeutet, dass diese Attribute ohne Generalisierung veröffentlicht werden. Ansonsten macht die Veröffentlichung keinen Sinn. Wenn sich die Vermutungen des Angreifer von *A0* auf *A1* stark verändern, kann davon ausgegangen werden, dass der Angreifer viele neue Informationen durch die veröffentlichten Daten gewonnen hat, wie zum Beispiel seine Vermutungen wurden widerlegt. Kurz gesagt, desto höher der Unterschied zwischen *A0* und *A1* ist, desto wertvoller sind die veröffentlichten sensiblen Informationen. Da der Informationsgewinn zwischen *A0* und *A1* jedoch die ganze Datenbank betrifft und keine Individuen oder Äquivalenzklassen, ist er zu vernachlässigen. Der Informationsgewinn zwischen *A1* und *A2* ist der zu beobachtende entscheidende Schritt. Der Informationsgewinn zwischen *A1* und *A2* kann klein gehalten werden, wenn der Unterschied zwischen *P* und *Q* niedrig gehalten wird. Kurz gesagt, wenn *P=Q* dann *A1=A2.* Desto ähnlicher die Verteilung der sensiblen Attribute einer Äquivalenzklasse der Gesamtverteilung der sensiblen Attribute des Datensatzes entspricht, desto geringer ist der Informationsgewinn des Angreifers zwischen A1 und A2. Diese Verteilung der sensiblen Attribute wird durch die *t-Closeness* Methode berechnet und beziffert [48, p. 109].

4.9 Angriffe auf die Anonymität

In diesem Kapitel werden einige Angriffe auf die bisher vorgestellten Anonymitätsverfahren erklärt. Die Angriffsarten können unterteilt werden in *Linking-Attacken* und *probabilistischen Attacken*. Erstere beschreibt Attacken in denen Merkmale oder Identitäten einem einzigen Dateneintrag zugeordnet werden können. Dabei kann zwischen *Attribute, Table und Record linking* unterschieden werden. Damit diese vermieden werden können, darf dem Angreifer so wenig Hintergrundwissen wie möglich zur Verfügung stehen [16, p. 34].

4.9.1 Unsorted Matching-Attack

Der *Unsorted Matching-Angriff* fällt in die Kategorie der *Linking-Attacken* und tritt auf, wenn die Reihenfolge der Datensätze mithilfe des Anonymisierungsvorgangs nicht verändert wird. Die Schwachstelle kann ausgenutzt werden, wenn wie in Abbildung 21 zwei Teile einer Datenbank veröffentlicht werden, die nicht miteinander verknüpft werden dürfen [16, p. 35].

Direkter Identifikator	Quasi-Identifikator
Vollständiger Name	Geschlecht
Josef Meier	M
Marie Holzer	W
Sarah Thompson	W
Tobias Rend	M
Ruth Talin	W
Sebastian Eisberger	M
Jasmin Scheuer	W
Gernot Flick	M
Friedrich Scheuer	M

Quasi-Identifikator		Sensibles Attribut
Alter	PLZ	Politische Gesinnung
20	1060	SPÖ
31	1020	SPÖ
22	1070	ÖVP
33	1030	ÖVP
44	1150	SPÖ
45	1160	KPÖ
26	1080	GRÜNE
37	1040	GRÜNE
48	1170	GRÜNE

Abbildung 21: Anfällig für *Unsorted Matching-Angriff* [16, p. 35]

In Abbildung 21 sind zwei Teile einer Datenbank, die unabhängig voneinander veröffentlicht worden sind, zu sehen. Dabei ist die Reihenfolge der Datensätze unverändert geblieben und die Verknüpfbarkeit somit ein

einfach durchzuführender Angriff. Es kann durch Aneinanderreihen der einzelnen Personen die politische Gesinnung jeder einzelnen Person wiederhergestellt werden.

4.9.2 Complementary release-Attack

Die *Complementary release* Attacke kann durchgeführt werden, wenn aus einer Datenbank zwei unabhängige Gruppen von Quasi-Identifikatoren als zwei eigenständige Datenbanken veröffentlicht werden und dabei unterschiedlich bei der Anonymisierung der Quasi-Identifikatoren beider Datenbanken vorgegangen wurde.

Quasi-Identifikatoren				Sensibles Attribut
Hautfarbe	Geburtsjahr	Geschlecht	ZIP	Krankheit
Schwarz	1980	M	02159	Asthma
Schwarz	1980	M	02159	Brustschmerzen
Schwarz	1980	W	02141	Augenschmerzen
Schwarz	1980	W	02141	Schnupfen
Schwarz	1979	W	02141	Übergewichtig
Schwarz	1979	W	02141	Brutschmerzen
Weiß	1979	M	02141	Kurzatmigkeit
Weiß	1980	W	02142	ADHS
Weiß	1979	M	02142	Übergewichtig
Weiß	1979	M	02142	Schnupfen
Weiß	1982	M	02141	Übelkeit
Weiß	1982	M	02141	Kopfschmerzen

Abbildung 22: Originaldatensatz [56, p. 11]

Abbildung 22 besteht aus dem Originaldatensatz mit den Quasi-Identifikatoren Hautfarbe, Geburtsjahr, Geschlecht, Zip Code und dem sensiblen Merkmal Krankheit. Im nächsten Schritt wird der Originaldatensatz in zwei verschieden anonymisierten Datenbanken veröffentlicht.

Quasi-Identifikatoren				Sensibles Attribut
Hautfarbe	Geburtsjahr	Geschlecht	ZIP	Krankheit
Schwarz	1980	M	02159	Asthma
Schwarz	1980	M	02159	Brustschmerzen
*	1980	W	0214*	Augenschmerzen
*	1980	W	0214*	Schnupfen
Schwarz	1979	W	02141	Übergewichtig
Schwarz	1979	W	02141	Brutschmerzen
Weiß	1979	M	0214*	Kurzatmigkeit
*	1980	W	0214*	ADHS
Weiß	1979	M	0214*	Übergewichtig
Weiß	1979	M	0214*	Schnupfen
Weiß	1982	M	02141	Übelkeit
Weiß	1982	M	02141	Kopfschmerzen

Abbildung 23: k=2 Anonymity, Datenbank1

Quasi-Identifikatoren				Sensibles Attribut
Hautfarbe	Geburtsjahr	Geschlecht	ZIP	Krankheit
Schwarz	1980	M	02159	Asthma
Schwarz	1980	M	02159	Brustschmerzen
Schwarz	1980	W	02141	Augenschmerzen
Schwarz	1980	W	02141	Schnupfen
Schwarz	1979	W	02141	Übergewichtig
Schwarz	1979	W	02141	Brutschmerzen
Weiß	1979-83	M	02141	Kurzatmigkeit
Weiß	1979-83	*	02142	ADHS
Weiß	1979-83	*	02142	Übergewichtig
Weiß	1979-83	*	02142	Schnupfen
Weiß	1979-83	M	02141	Übelkeit
Weiß	1979-83	M	02141	Kopfschmerzen

Abbildung 24: k=2 Anonymity, Datenbank2

Durch die *Linking Attacke* wurden die zwei unterschiedlich anonymisierten und veröffentlichten Datenbanken über das sensible Attribut Krankheit verknüpft. Dadurch konnte der Originaldatensatz und somit der Personenbezug wiederhergestellt werden. Die Person mit ADHS konnte somit als weiblich identifiziert werden. Außerdem weiß der Angreifer durch die Attacke, dass die Person mit ADHS, weiß ist und 02142 als ZIP-Code eingetragen hat [56, pp. 11-12].

4.9.3 Temporal Attack

Bei der *Temporal Attack* [56, p. 12] werden alte veröffentliche Versionen einer Datenbank als Hintergrundwissen verwendet. Dadurch können Personen in der neu veröffentlichen Version einer Datenbank durch *Linking Attacken* Re-Identifiziert werden. Als Gegenmaßnahmen sollten alle bereits veröffentlichten Versionen evaluieren und dokumentieren, ob ein Personenbezug durch Kombination wiederhergestellt werden kann.

4.9.4 Homogenety Attack

Die *Homogenety Attack* [55] kann durchgeführt werden, wenn bei k-anonymen Äquivalenzklassen nicht auf die Diversität der sensiblen Attribute geachtet wird. Angenommen Alice und Bob sind Nachbarn, die in derselben Straße wohnen. Eines Tages wird Bob plötzlich schwer krank und wird vom Krankenwagen ins Krankenhaus gebracht. Alice ist eine sehr neugierige Person und möchte herausfinden an welcher Krankheit Bob leidet. Da sie seine Nachbarin ist, weiß sie sie Postleitzahl, das Alter und das Geschlecht. Sie gelangt also an veröffentlichte anonyme Krankenhausdaten, in denen sich Bob befindet. Die anonyme Datenbank ist k=4 anonym und jede Äquivalenzklasse beinhaltet mindestens 4 idente Kombinationen aus Quasi-Identifikatoren. Dadurch kann Alice die Äquivalenzklasse, in der sich Bob befindet ausfindig machen und das sensible Attribut Krankheit auslesen. Da nun alle Einträge der Äquivalenzklasse von Bob an HIV erkrankt sind, kann Alice auf die Krankheit von Bob durch eine *Homogenety Attack* rückschließen. In Kapitel 4.8.1, Abbildung 17 ist ein praktisches Beispiel angeführt.

4.9.5 Background Knowledge Attack

Bei der Anonymisierung von Daten kann während des Prozesses evaluiert werden, welches Hintergrundwissen der Angreifer bereits besitzen könnte. Dieses Hintergrundwissen kann für die *Background Knowledge-Attack* verwendet werden und führt zu Re-Identifizierung von Subjekten einer Datenbank. Der Angreifer kann durch Hintergrundwissen, welches aus öffentlichen Quellen stammt oder allgemein bekannt ist mit gewisser Wahrscheinlichkeit einzelne Individuen in einem Datensatz ausmachen und ihre Attribute identifizieren [50, p. 14].

4.9.6 Similarity Attack

Bei der *Similarity Attack* [48, p. 108] kann trotz l-Diversität die Aussagekraft des sensiblen Attributs eines gesuchten Subjekts in der Datenbank bestimmt werden. Die Schwachstelle tritt auf, wenn zwar die Bezeichnungen des sensiblen Attributs einer Äquivalenzklasse divers sind, jedoch ihre Bedeutung die gleiche ist. Dadurch kann ein Angreifer zum Beispiel die ungefähre Richtung der politischen Gesinnung einer gesuchten Person bestimmen. In Kapitel 4.8.2, Abbildung 20 ist ein praktisches Beispiel angeführt.

4.9.7 Skewness Attack

Bei der *Skewness Attacke* wird die mögliche ungleiche Verteilung von sensiblen Merkmalswerten trotz Erfüllung der l-Diversität ausgenutzt. Angenommen eine Äquivalenzklasse hat 100 Einträge und hat das Kriterium der 2-Diversität zu Erfüllen. Nun ist dieser Maßstab bereits erfüllt, wenn 99 der Einträge als sensiblen Merkmalswert *positiv* und 1 Eintrag den sensiblen Merkmalswert *negativ* innehat. Ein Individuum dieser Äquivalenzklasse wäre also zu 99% *positiv*. Dadurch kann der Angreifer mit 99-prozentiger Wahrscheinlichkeit vorhersagen, dass sein gesuchtes Subjekt in dieser Äquivalenzklasse als sensiblen Merkmalswert *positiv* besitzt.

5 Experiment

Das vorliegende Experiment dient zur Beantwortung der Forschungsfrage und zum besseren praktischen Verständnis wie eine Re-Identifizierung eines Subjekts in einer Datenbank erfolgen kann. Das Kapitel verwendet einen öffentlich zugänglichen Datensatz und leicht verständliche Ansätzen zur Re-Identifizierung. Damit den einzelnen Schritten des Experiments leicht gefolgt werden kann, werden die Gegebenheiten und Grundlagen des Datenbestandes im Voraus erklärt.

5.1 Verwendete Programme und Bibliotheken

Zur Programmierung und Darstellung der Code Ausschnitte wurde die Webanwendung *Jupyter Notebook* verwendet. *Jupyter Notebook* [57] ist seit 2014 ein *open source* Projekt und ist aus dem *IPython Projekt* entstanden. Das Programm dient zur Erstellung und zum Teilen von Programmen. Außerdem bietet *Jupyter Notebook* eine übersichtliche Benutzeroberfläche, die für ein strukturiertes und einfaches Programmiererlebnis sorgt. Der große Vorteil dabei ist, dass jede Codezeile einzeln in sogenannten *Codeblöcken* ausgeführt werden kann und dann das Ergebnis sofort sichtbar ist. Dabei ist man als Programmierer selbst an keine Programmiersprache gebunden und kann über 40 verschiedene Programmiersprachen verwenden. *Jupyter Notebook* wird über die Konsole gestartet mit dem Befehl *Jupyter Notebook.* Daraufhin öffnet sich die Weboberfläche und das gewünschte Programm kann geöffnet werden. Die Kommentarzeilen können standardmäßig mit einem Rautezeichen markiert werden, damit *Jupyter Notebook* diese nicht ausführt und sie als Kommentar wertet. Das dient der besseren Nachvollziehbarkeit und kann zu Dokumentationszwecken verwendet werden. Die Daten werden in *Jupyter Notebook* durch verschiedene Programmbibliotheken aufbereitet, um in weiterer Folge verarbeitet werden zu können. In diesem Experiment wurde die Programmiersprache *Python* inklusive der *pandas* Bibliothek verwendet.

Pandas [58] ist seit 2009 eine *open source* Bibliothek und wird von einer freiwilligen Community weiterentwickelt. Dabei ist es möglich durch *pandas* gewisse *Dataframes* zu erzeugen, in denen die zu verarbeitenden Daten aufbereitet werden können. Zusätzlich bietet *pandas* die Möglichkeit externe Datenfiles zu importieren und mit den importierten Daten weiterzuarbeiten. Dabei gibt es verschiedene Möglichkeiten die importierten Datenstrukturen zu verändern und fehlende Werte sowie Duplikate zu behandeln. Durch die externe Programmbibliothek können dem *Dataframe* Spalten hinzugefügt aber auch Spalten weggenommen werden. Dabei sind dem Programm keine Grenzen gesetzt, um den unterschiedlichen Anforderungen gerecht zu werden. Wenn sich dann die Daten in der richtigen Datenstruktur befinden, können Gruppen gebildet, Graphen erstellt und Rechenoperationen durchgeführt werden. *Pandas* stellt hier viele Werkzeuge zur Verfügung, um aus den Daten den gewünschten Output zu bekommen.

5.2 Verwendeter Datensatz

Um die Forschungsfrage dieser Diplomarbeit zu beantworten, wurde ein öffentlich zugänglicher Datensatz, welcher Taximeterdaten von 442 Taxis aus Porto in Portugal zwischen dem 01.07.2013 bis zum 30.06.2014 enthält und von der Forschungswebsite „Kaggle“ [11] stammt, verwendet. Anhand dieser Daten wurde ein vertiefendes Experiment durchgeführt. Jedes Taxi in diesem Datensatz, hat zur Koordination und zur Aufzeichnung ein mobiles Datenterminal im Auto montiert. Darüber werden die Taxis von der Taxizentrale aus koordiniert und ihre Fahrten werden über das GPS-System aufgezeichnet. Bei einer aktiven Fahrt zeichnet dieses mobile Terminal die Start- sowie Endkoordinaten und alle 15 Sekunden die GPS-Lokation des Taxis in einem Protokoll auf. Dabei handelt es sich um einen pseudonymisiert Datensatz, der keinen direkten Personenbezug zulässt.

Die Datenbank besteht in ihrem Originalformat aus 9 Spalten: [11]

1) TRIP_ID: Beinhaltet pro Fahrt eine einzigartige TRIP_ID
2) CALL_TYPE: Beinhaltet die Information wie die Fahrt angefordert wurde
    a. A) Wenn die Fahrt von der Zentrale in Auftrag gegeben wurde.
    b. B) Wenn die Fahrt direkt am Taxistand durch eine Person in Auftrag gegeben wurde.
    c. C) Alles andere, wie zum Beispiel in einer zufälligen Straße einen Kunden angenommen.
3) ORIGIN_CALL: Beinhaltet eine einen einzigartigen Identifikator für jede Telefonnummer, die eine Fahrt beantragt hat und als CALL_TYPE = A eingetragen hat. Ansonsten enthält das Merkmal einen NULL Wert.
4) TAXI_ID: Beinhaltet eine einzigartige TAXI_ID für jeden Taxifahrer, der eine Fahrt gefahren ist.
5) TIMESTAMP: Beinhaltet die Startuhrzeit der Fahrt.
6) DAYTPE: Bezeichnet den Tag an dem die Fahrt stattgefunden hat und kann auf drei Kategorien aufgeteilt werden:
    a. B) Wenn die Fahrt an einem Feiertag oder sonstigen Spezialtagen stattfindet.
    b. C) Wenn die Fahrt an einem Tag bevor eines B-Typ Tages stattfindet.
    c. A) Alle anderen Tage (normaler Wochentag, Wochenende)
7) MISSING_DATA: Beinhaltet einen *FALSE* Wert, wenn das GPS-Protokoll der Fahrt vollständig ist und *TRUE,* wenn im GPS-Protokoll eine oder mehrere GPS-Lokationen fehlen dann *FALSE.*
8) POLYLINE: Beinhaltet alle GPS-Lokationsdaten pro Fahrt im WGS84 Format. Dabei werden für jede Fahrt die Start- und Endkoordinaten gesammelt und in einer Liste pro Fahrt zur Verfügung gestellt. Während der Fahrt werden alle 15 Sekunden die aktuellen GPS-Daten des Fahrzeuges in die Liste mit aufgenommen und protokolliert.
   Das Format der GPS-Daten ist *[LONGITUDE, LATITUDE].*

Das WGS84 *(World Geodetic System)* ist ein internationales geodätisches Referenzsystem, welches als einheitliche Grundlage zur Darstellung von Positionsangaben verwendet wird und sich deshalb zur Darstellung von GPS-Daten eignet [59]. Damit während der Durchführung des Experiments keine unnötigen Daten aufscheinen, wurden nur die wichtigen Merkmale des Datensatzes aufbereitet und bearbeitet. Alle unwichtigen Merkmale sind in den ersten Schritten entfernt worden.

5.3 Ziel des Experiments

Das Ziel des Experiments ist es aufzuzeigen, dass obwohl die Kriterien einer k=8 Anonymität erreicht wurden und der Datensatz gängigen Anonymisierungsverfahren unterzogen wurde, allein das Wissen über eine dritte GPS-Koordinate einer Fahrt ausreicht, um auf ein explizites Taxi rückschließen zu können.

5.4 Durchführung des Experiments

```
# Praktisches Beispiel für Diplomarbeit 2022 - Hinterholzer Matthias is201851 FH. St. Pölten - IT Security
# Importieren von notwendigen Libraries und Deaktivieren von unnötigen Warnmeldungen
import pandas as pd
pd.options.mode.chained_assignment = None  # default='warn'
```

```
# Daten von train.csv werden in Dataframe eingelesen
dffirst = pd.read_csv('train.csv', delimiter=',')
dffirst.shape
```

```
(1710670, 9)
```

```
# Ersten Zeilen des Dataframes inkl. Spaltennamen werden angezeigt
dffirst.head()
```

	TRIP_ID	CALL_TYPE	ORIGIN_CALL	ORIGIN_STAND	TAXI_ID	TIMESTAMP	DAY_TYPE	MISSING_DATA	POLYLINE
0	1372636858620000589	C	NaN	NaN	20000589	1372636858	A	False	[[-8.618643,41.141412], [-8.618499,41.141376],[...
1	1372637303620000596	B	NaN	7.0	20000596	1372637303	A	False	[[-8.639847,41.159826], [-8.640351,41.159871],[...
2	1372636951620000320	C	NaN	NaN	20000320	1372636951	A	False	[[-8.612964,41.140359], [-8.613378,41.14035],[-...
3	1372636854620000520	C	NaN	NaN	20000520	1372636854	A	False	[[-8.574678,41.151951], [-8.574705,41.151942],[...
4	1372637091620000337	C	NaN	NaN	20000337	1372637091	A	False	[[-8.645994,41.18049], [-8.645949,41.180517],[-...

Abbildung 25: Schritt 1

In den ersten Schritten werden die *pandas* Datenbank importiert und unwichtige Warnnachrichten deaktiviert. Danach wird die Datenbank mit den Testdaten importiert. Mit dem Befehl *read_csv* wird die CSV Datei eingelesen und das Trennzeichen bekannt gegeben. Der erste *Dataframe* wird *dffirst* getauft und gleich dazu verwendet, um mit der *head* Funktion die Zeilen und Spalten des Datensatzes auszugeben.

```
# Nicht benötigte Spalten werden gedroppt
dffirst = dffirst.drop(["CALL_TYPE", "ORIGIN_CALL","ORIGIN_STAND","DAY_TYPE", "MISSING_DATA", 'TIMESTAMP'], axis=1)
```

```
# Konvertieren von unix epoch zu human readable time (nicht verwendet)
# df['TIMESTAMP'] = pd.to_datetime(df["TIMESTAMP"],unit="s", utc="true")
# df['TIMESTAMP'] = df['TIMESTAMP'].dt.tz_convert('Europe/Berlin')
# Neu Benennung der Spalte Polyline zu Coordinates
dffirst.columns = dffirst.columns.str.replace("POLYLINE","COORDINATES")
# Regex die Start- und Endkoordinaten aus den GPS-Protokollen
dffirst["LatitudeStartpoint"] = dffirst["COORDINATES"].str.extract(r"\[\[[\+-]\d+\.\d+\s*,\s*(?P<LatitudeStartpoint>\d+
dffirst["LongitudeStartpoint"] = dffirst["COORDINATES"].str.extract(r"\[\[(?P<LongitudeStartpoint>[\+-]\d+\.\d+)\s*,\s*'
dffirst["LatitudeEndpoint"] = dffirst["COORDINATES"].str.extract(r"\[[\+-]\d+\.\d+\s*,\s*(?P<LatitudeEndpoint>\d+\.\d+)'
dffirst["LongitudeEndpoint"] = dffirst["COORDINATES"].str.extract(r"\[(?P<LongitudeEndpoint>[\+-]\d+\.\d+)\s*,\s*\d+\.\(

```

```
# Drop von TripIDs ohne Koordinaten und anschließende Ausgabe wieviele Zeilen und Spalten der Dataframe noch hat
dffirst.dropna(subset = ["LatitudeStartpoint"], inplace=True)
dffirst.shape
```

```
(1704768, 7)
```

Abbildung 26: Schritt 2

Im zweiten Schritt werden die unwichtigen Tabellen entfernt und die Spalte mit den GPS-Protokolle der einzelnen Fahrten in *COORDINATES* umbenannt. Im weiteren Verlauf wird durch *dropna* die Spalte *LatitudeStartpoint* von allen *NaN* Werten bereinigt, um danach die Datensatzgröße mit dem Befehl *shape* nochmals auszugeben. Dabei ist eine Verkleinerung des Datensatzes von Abbildung 25 zu Abbildung 26 zu beobachten.

```
# Duplicate Rows werden erkannt und herausgefilter
duplicateRowsDF = dffirst[dffirst.duplicated(['LatitudeStartpoint','LongitudeStartpoint','LatitudeEndpoint',
                                              'LongitudeEndpoint'],keep=False)]
```

```
# Die Koordinaten Liste wird von den eckigen Klammern befreit
duplicateRowsDF["coordinat"]=duplicateRowsDF["COORDINATES"].apply(lambda x: x.replace("[[","").replace("]]","")
                                                                .replace("],[",";").split(";"))

# Eine neue Stopp Spalte wird erstellt mit der Anzahl der gesamten Koordinaten-Punkten einer Fahrt
duplicateRowsDF["stops"] = duplicateRowsDF.coordinat.apply(lambda x: len(x))

# Nur die Fahrten die mehr als 2 Stops in ihren GPS-Protokollen aufgezeichnet haben werden weiterverwendet
dfnew = duplicateRowsDF[duplicateRowsDF['stops'].apply(lambda x: x > 2)]
dfnew
```

			[-8.648037,41.170599],[...					
15136	1372929672620000018	20000018	[[-8.648289,41.170419], [-8.648037,41.170599],[...	41.170419	-8.648289	41.169339	-8.645022	[-8.648289,41 -8.648037,41
15139	1372929710620000018	20000018	[[-8.648289,41.170419], [-8.648037,41.170599],[...	41.170419	-8.648289	41.169339	-8.645022	[-8.648289,41 -8.648037,41
15142	1372929719620000018	20000018	[[-8.648289,41.170419], [-8.648037,41.170599],[...	41.170419	-8.648289	41.169339	-8.645022	[-8.648289,41 -8.648037,41
15144	1372929729620000018	20000018	[[-8.648289,41.170419], [-8.648037,41.170599],[...	41.170419	-8.648289	41.169339	-8.645022	[-8.648289,41 -8.648037,41
15145	1372929739620000018	20000018	[[-8.648289,41.170419], [-8.648037,41.170599],[...	41.170419	-8.648289	41.169339	-8.645022	[-8.648289,41 -8.648037,41

Abbildung 27: Schritt 3

In Schritt 3 werden alle Fahrten, deren Koordinaten der Merkmale *LatitudeStartpoint, LongitudeStartpoint, LatitudeEndpoint, LongitudeEndpoint* mit mindestens einer weiteren Fahrt ident sind, in einen neuen *Dataframe duplicateRowsDF* gespeichert. Dieser Vorgang wird durch den Befehl *duplicated* durchgeführt. Im zweiten Codeblock der Abbildung 27 wird die Koordinatenliste von den eckigen Klammern befreit, um so eine besser lesbare Koordinatenliste namens *coordinat* zu erstellen. Danach werden die Einträge der neu erstellten Liste gezählt und daraus Spalte *stops* abgeleitet, um so die Anzahl der GPS-Koordinatenpunkte pro Fahrt anzeigen zu können. Die Anzahl der *stops* wird im weiteren Verlauf zur Filterung und zur Erstellung des *Dataframes dfnew* verwendet. Dabei sind für den weiteren Verlauf des Experiments ausschließlich Fahrten mit mehr als zwei Stopps interessant.

```
# Gesamt dataframe Anzahl von Spalten und Reihen
duplicateRowsDF.shape
```

```
(2303, 9)
```

```
# Rows mit weniger als 2 Koordinaten-Punkten in einer Fahrt
sum(duplicateRowsDF['stops'].apply(lambda x: x <= 2))
```

```
1993
```

```
# Neuer Dataframe Anzahl an Fahrten mit mehr als 2 Koordinaten-Punkten
dfnew.shape
```

```
(310, 9)
```

Abbildung 28: Schritt 4

In Schritt 4 werden die *Dataframes* überprüft ob bei der Filterung der Fahrten auch keine Einträge verloren gegangen sind. Dabei ist zu erkennen, dass nach der Filterung auf Fahrten mit mehr als zwei Stopps sowie auf Duplikate bezüglich Start- und Endkoordinaten der Datensatz erheblich geschrumpft ist. 2303 Fahrten haben mindestens eine weiteren Fahrt, die idente Start- und Endkoordinaten besitzt. Davon besitzen lediglich 310 Fahrten mehr als zwei Stopps. Die restlichen 1993 Fahrten sind für den weiteren Verlauf des Experiments uninteressant.

```
# Anzeige von Fahrten wo es eine zweite mit identen start und endkoordinatenpunkt gibt
dfnewduplicated = dfnew[dfnew.duplicated(['LatitudeStartpoint','LongitudeStartpoint','LatitudeEndpoint',
                                          'LongitudeEndpoint'],keep=False)]
```

```
# Erzeugt eine Spalte um pro Koordinaten Gruppe die Anzahl von Uniquen TaxiIDS anzuzeigen
dfnewduplicated['TaxiIDCOUNT'] = dfnewduplicated.groupby(['LatitudeStartpoint', 'LongitudeStartpoint',
                                 'LatitudeEndpoint', 'LongitudeEndpoint'])['TAXI_ID'].transform('nunique').astype(int)
```

```
# Nur die Gruppen mit mehr als 1 einzigartigen TaxiIDs werden weiterverwendet
dspecial = dfnewduplicated[dfnewduplicated['TaxiIDCOUNT'].apply(lambda x: x > 1)]
```

```
dspecial
```

25020	1373073592620000297	20000297	[[-8.612541,41.145993], [-8.612316,41.146038],[...	41.145993	-8.612541	41.155614	-8.602614	[-8.612541,41 -8.612316,41
35406	1373270938620000027	20000027	[[-8.606502,41.144679], [-8.607213,41.144643],[...	41.144679	-8.606502	41.148882	-8.585595	[-8.606502,41 -8.607213,41
38678	1373310086620000604	20000604	[[-8.607753,41.15259], [-8.607771,41.152581],[-...	41.15259	-8.607753	41.162301	-8.660556	[-8.607753,4 -8.607771,41
38939	1373315455620000671	20000671	[[-8.615538,41.140683], [-8.615169,41.140845],[...	41.140683	-8.615538	41.146812	-8.620038	[-8.615538,41 -8.615169,41
49352	1373524100620000669	20000669	[[-8.610867,41.145669], [-8.610858,41.145678],[...	41.145669	-8.610867	41.147217	-8.622405	[-8.610867,41 -8.610858,41
50794	1373533448620000051	20000051	[[-8.619849,41.147991], [-8.61984,41.147991],[-	41.147991	-8.619849	41.151609	-8.609472	[-8.619849,41 -8.61984,41

Abbildung 29: Schritt 5

Da sich nun die Anzahl der interessanten Fahrten verändert hat, müssen im ersten Codeblock erneut die Duplikate gefiltert werden. Die Fahrten mit Duplikaten sind für dieses Experiment die einzig interessanten Fahrten. Dabei wird ein neuer *Dataframe dfnewduplicated* erstellt. Durch die *groupby* Funktion wird im zweiten Codeblock die neue Spalte *TaxiIDCOUNT* erstellt. Dabei wird für jede Gruppe von Fahrten, welche idente Merkmalswerte in *LatitudeStartpoint, LongitudeStartpoint, LatitudeEndpoint, LongitudeEndpoint* besitzen, die einzigartig vorkommenden TaxiIDs in der Gruppe gezählt und neben jedem Eintrag angezeigt. Das bedeutet für alle Fahrten, deren Koordinaten der Merkmale *LatitudeStartpoint, LongitudeStartpoint, LatitudeEndpoint, LongitudeEndpoint* mit mindestens einer weiteren Fahrt ident sind, wird für jede Gruppe ausgegeben wie viele unterschiedliche Taxis vorkommen. Bei Gruppierung der Koordinaten auf die fünfte und sechste Kommastelle genau ergaben sich lediglich Gruppen mit höchstens zwei unterschiedlichen TaxiIDs.

```
# Die Koordinaten werden auf 3 Nachkommastellen gerundet und neue Gruppen werden aufgrunddessen gebildet
dspecial['LongitudeEndpoint']= dspecial['LongitudeEndpoint'].astype(float).round(3)
dspecial['LongitudeStartpoint']= dspecial['LongitudeStartpoint'].astype(float).round(3)
dspecial['LatitudeStartpoint']= dspecial['LatitudeStartpoint'].astype(float).round(3)
dspecial['LatitudeEndpoint']= dspecial['LatitudeEndpoint'].astype(float).round(3)
```

```
# Für jede gebildete Gruppe von Fahrten mit gleichen Start und Endkoordinaten
# werden die einzigartigen TaxiIDs pro Gruppe gezählt
dspecial['TaxiIDCOUNT'] = dspecial.groupby(['LatitudeStartpoint', 'LongitudeStartpoint','LatitudeEndpoint',
                                            'LongitudeEndpoint'])['TAXI_ID'].transform('nunique').astype(int)
```

```
# Nur die Gruppen mit mehr als 5 einzigartigen TaxiIDs werden weiterverwendet
dspecialnew = dspecial[dspecial['TaxiIDCOUNT'].apply(lambda x: x > 5)]
```

```
dspecialnew
```

	TRIP_ID	TAXI_ID	COORDINATES	LatitudeStartpoint	LongitudeStartpoint	LatitudeEndpoint	LongitudeEndpoint	coor
569278	1383220690620000617	20000617	[[-8.599248,41.149161], [-8.598762,41.148855],[...	41.149	-8.599	41.149	-8.586	[-8.599248,41.149 -8.598762,41.148
698751	1385633775620000054	20000054	[[-8.599257,41.149152], [-8.599239,41.149161],[...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.599239,41.149
842078	1388221685620000272	20000272	[[-8.599257,41.149152], [-8.599203,41.149143],[...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.599203,41.149
951388	1390422203620000256	20000256	[[-8.599257,41.149152], [-8.598618,41.148567],[...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.598618,41.148
985237	1391153533620000174	20000174	[[-8.599248,41.149161], [-8.598717,41.148729],[...	41.149	-8.599	41.149	-8.586	[-8.599248,41.149 -8.598717,41.148
1225201	1395682884620000376	20000376	[[-8.599257,41.149161], [-8.599257,41.14917],[-...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.599257,41.14 -
1306179	1397209892620000367	20000367	[[-8.599257,41.149161], [-8.598888,41.148972],[...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.598888,41.148
1349139	1397991969620000675	20000675	[[-8.599257,41.149152], [-8.599221,41.149161],[...	41.149	-8.599	41.149	-8.586	[-8.599257,41.149 -8.599221,41.149

Abbildung 30: Schritt 6

Damit diesem Problem entgegengewirkt werden kann, werden die Start- und Endkoordinaten zur besseren Gruppenbildung auf drei Kommastellen gerundet. Die Gruppen müssen mindestens sechs Fahrten mit identen Start- und Endkoordinaten beinhalten. Zusätzlich muss jede der Fahrten einer Gruppe eine einzigartige TaxiID besitzen. Daraufhin wurde mit der *groupby* Funktion die einzigartig vorkommenden TaxiIDs, je Gruppe gezählt und in der Spalte *TaxiIDCOUNT* notiert. Im nächsten Schritt wurde der neue *Dataframe dspecialnew* erstellt und ausschließlich mit Gruppen von Fahrten gefüllt, die mehr als fünf Fahrten mit identen Start- und Endkoordinaten sowie einzigartige TaxiIDs besitzen.

Das Ergebnis ist eine Gruppe von Fahrten, welche die k=8 Anonymitätskriterien erfüllt und somit durch ein gängiges Anonymitätsverfahren *(Rounding)* anonymisiert wurde. Durch das Wissen über Start- und Endkoordinaten sind die einzelnen Fahrten nicht mehr identifizierbar.

5.5 Ergebnis

```
# Nicht mehr verwendete Spalten werden gedroppt
dspecialnew.drop(['COORDINATES', 'TaxiIDCOUNT' ], axis=1, inplace=True)
```

```
# Die Gesuchte Koordinate wird in einer neue Spalte eingefügt
dspecialnew["InvestigationPoint"] = dspecialnew.coordinat.apply(lambda x: x[1])
dspecialnew
```

	TRIP_ID	TAXI_ID	LatitudeStartpoint	LongitudeStartpoint	LatitudeEndpoint	LongitudeEndpoint	coordinat	stops	Investiga
569278	1383220690620000617	20000617	41.149	-8.599	41.149	-8.586	[-8.599248,41.149161, -8.598762,41.148855, -8....	18	-8.598762,4
698751	1385633775620000054	20000054	41.149	-8.599	41.149	-8.586	[-8.599257,41.149152, -8.599239,41.149161, -8....	21	-8.599239,4
842078	1388221685620000272	20000272	41.149	-8.599	41.149	-8.586	[-8.599257,41.149152, -8.599203,41.149143, -8....	15	-8.599203,4
951388	1390422203620000256	20000256	41.149	-8.599	41.149	-8.586	[-8.599257,41.149152, -8.598618,41.148567, -8....	13	-8.598618,4
985237	1391153533620000174	20000174	41.149	-8.599	41.149	-8.586	[-8.599248,41.149161, -8.598717,41.148729, -8....	16	-8.598717,4
1225201	1395682884620000376	20000376	41.149	-8.599	41.149	-8.586	[-8.599257,41.149161, -8.599257,41.14917, -8.5...	25	-8.599257,-
1306179	1397209892620000367	20000367	41.149	-8.599	41.149	-8.586	[-8.599257,41.149161, -8.598888,41.148972, -8....	15	-8.598888,4
1349139	1397991969620000675	20000675	41.149	-8.599	41.149	-8.586	[-8.599257,41.149152, -8.599221,41.149161, -8....	18	-8.599221,4

Abbildung 31: Ergebnis

In Abbildung 31 ist zu sehen, dass die Spalten *COORDINATES* und *TaxiIDCOUNT* nicht mehr benötigt und daher gedroppt werden. Im zweiten Codeblock wird davon ausgegangen, dass der Angreifer neben den Start- und Endpunkten der einzelnen Fahrten, einen dritten Punkt des GPS-Protokolls der einzelnen Fahrten besitzt. Dieser Punkt wird in der Tabelle als *InvestigationPoint* bezeichnet und soll der Re-Identifizierung einzelner Taxis trotz k=8 Anonymität dienen.

```
# Es wird nach einer Koordinate in den GPS-Protokollen der Fahrten der Gruppe gesucht
dspecialnew['coordinat'].apply(lambda x: "-8.598762,41.148855" in x)
```

```
569278     True
698751    False
842078    False
951388    False
985237    False
1225201   False
1306179   False
1349139   False
Name: coordinat, dtype: bool
```

Abbildung 32: Ergebnis

Abbildung 32 zeigt den erfolgreichen Versuch der Re-Identifizierung. Dabei wird das GPS-Protokoll jeder Fahrt nach der dritten Koordinate durchsucht. Wenn die dritte Koordinate in einem GPS-Protokoll der jeweiligen Fahrt vorkommt, so wird die Fahrt als TRUE markiert.

Hierbei wird erkenntlich, dass sobald der Angreifer einen dritten Punkt neben Start- und Endkoordinate des GPS-Protokolls weiß (in diesem Fall eine dritte GPS-Koordinate der ersten Fahrt), kann er trotz erfüllter k=8 Anonymitätskriterien und anonymisierter Datenbank auf ein einziges Taxi rückschließen.

6 Conclusio

Das Ziel des letzten Kapitels ist es, die Erkenntnisse der Arbeit zusammenfassend wiederzugeben. Dabei werden die vorhandenen Ergebnisse kurz und bündig vorgestellt.

Die stetig wachsenden Datenströme des digitalen Zeitalters stellen eine Bedrohung für die Privatsphäre des Individuums dar. Bei 4.94 Milliarden [7] aktiven Internetnutzern und einer voraussichtlich generierten Datenmenge von 175 Zettabytes [1] im Jahr 2025, spielen Datenschutz und die dazugehörigen Anonymisierungsverfahren eine wesentliche Rolle. Dabei zeigten die Ergebnisse des Datenschutz Kapitels, dass bei einer Gruppe von 1137 Personen älter als 14 Jahren 30% davon keine und 25% eher selten Datenschutzbestimmungen lesen [17]. Die Hälfte der Befragten wissen also kaum bzw. nicht was mit ihren personenbezogenen Daten geschieht und wie sie verarbeitet werden. In einer Zeit, in der Privatpersonen personenbezogene Daten in den sozialen Medien freiwillig veröffentlichen, ist das eine kritische Anzahl an Unwissenden. Damit dem entgegengewirkt werden kann, wurden im Datenschutzkapitel drei Säulen [16, p. 4] erläutert. Die erste Säule beschäftigt sich mit den gesetzlichen Regulatorien die ebenfalls in dieser Arbeit erläutert wurden. Die zweite Säule beschäftigt sich mit Sicherheitsmaßnahmen, die ein Unternehmen zum Schutz seiner Kunden treffen kann. Die dritte und letzte Säule beschreibt die Schutzmaßnahmen, die durch die Personen selbst getroffen werden können.

Das rechtliche Kapitel dieser Arbeit zeigte, dass die DSGVO [10] die Rechte der europäischen Bürger umfassend bestärkt. Gleichzeitig werden die Verantwortlichen zu Schutzmaßnahmen und der korrekten Datenverarbeitung verpflichtet. Das Kapitel bringt zusätzlich zum Vorschein, dass im Vergleich zu den verschiedenen Bundes- und Landesgesetzen für einzelne Sektoren der USA, die DSGVO [10] einen umfassenden Schutz der personenbezogenen Daten für europäische Bürger bietet. Aus dem Kapitel geht ebenfalls hervor, dass personenbezogene Daten ausreichend anonymisiert sind, wenn sie den Status der faktischen Anonymisierung erreicht haben. Faktische Anonymisierung beschreibt den Zustand von Daten, in denen sie ausreichend anonymisiert sind und nicht mehr in den Rechtsbereich der DSGVO [10] fallen. Dabei können sie trotz Anonymisierung für statistische Auswertungen oder Studien verwendet werden. Laut Definition ist es dem Angreifer im Fall der faktischen Anonymisierung „nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft“ möglich, einzelne Merkmale dem Merkmalsträger zuzuordnen.

Das Kapitel der Anonymisierung zeigt, dass es zwei Arten von Anonymisierungsverfahren gibt. Die *Datenverändernden Methoden* und die *Datenaggregierenden Methoden*. Dabei ist die *k-Anonymity* als Anonymisierungskriterium ein weit verbreiteter Maßstab zur Messung der Anonymität eines Datensatzes. Dabei konnte gezeigt werden, wie die *k-Anonymity* durch die *l-Diversity* und *t-Closeness* erweitert werden und somit die Personen eines Datensatzes vor *Identity Disclosure* geschützt sind.

Damit die Forschungsfrage durch diese Arbeit beantwortet wird, ist ein Experiment mit realen, aus der EU stammenden Bewegungsdaten durchgeführt worden. Dabei wurden die Daten so aufbereitet, dass eine Gruppe von acht Subjekten mit jeweils den gleichen Start- und Endkoordinaten gebildet werden konnte. Durch das Experiment wurde gezeigt, dass sobald ein Angreifer zusätzlich zu den veröffentlichten Start- und Endkoordinaten über eine dritte GPS-Koordinate seines Ziels Bescheid weiß, dieses eindeutig in der Gruppe identifiziert werden kann. Daraus lässt sich schließen, dass es in der heutigen Zeit, in der Personen leichtfertig ihren Standort über soziale Medien preisgeben möglich ist, gesuchte Personen in DSGVO [10] konform anonymisierten Datensätzen mit einfachen Mitteln zu identifizieren.

1. Literaturverzeichnis

[1] F. Tenzer, „de.statista.com," 24 01 2022. [Online]. Available: https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/. [Zugriff am 16 04 2022].

[2] T. Niesen, C. Houy, P. Fettke und P. Loos, „Towards an Integrative Big Data Analysis Framework for Data-driven Risk Management in Industry 4.0," 49th Hawaii International Conference on System Sciences, 2016.

[3] „Plattform Industrie 4.0," [Online]. Available: https://www.plattform-i40.de/IP/Navigation/DE/Industrie40/WasIndustrie40/was-ist-industrie-40.html. [Zugriff am 16 04 2022].

[4] P. Bajpai, „investopedia," 20 09 2021. [Online]. Available: https://www.investopedia.com/articles/investing/030916/how-uber-uses-its-data-bank.asp. [Zugriff am 16 04 2022].

[5] A. Narayanan und V. Shmatikov, „Robust De-anonymization of Large Sparse Datasets," IEEE Symposium on Security and Privacy, Austin, 2008.

[6] V. Torra, „Data Privacy: Foundations, New Developments and the Big Data Challenge," Springer, 2017.

[7] S. Kemp, „datareportal," 26 01 2022. [Online]. Available: https://datareportal.com/reports/digital-2022-local-country-headlines. [Zugriff am 16 04 2022].

[8] A. G. C. &. Specialty, „Allianz Risk Barometer," Allianz Global Corporate & Specialty, 2022.

[9] wko, „wko," 01 04 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung.html. [Zugriff am 16 04 2022].

[10] EU, „VERORDNUNG (EU) 2016/679 DES EUROPÄISCHEN PARLAMENTS UND DES RATES," EU, 2016.

[11] kaggle, „kaggle," kaggle, [Online]. Available: https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data. [Zugriff am 17 04 2022].

[12] R. Petric, „Das vermessene Selbst," Datenschutz und Datensicherheit DuD, 2016.

[13] K. A. -. D. 18/9058, „deutscher bundestag," 04 07 2016. [Online]. Available: https://dserver.bundestag.de/btd/18/090/1809058.pdf. [Zugriff am 21 04 2022].

[14] C. Just und J. Struck, „computerbild," 18 06 2021. [Online]. Available: https://www.computerbild.de/artikel/cb-Tests-Sport-Wearables-Sportuhr-Smartwatch-Fitnessarmband-Test-9124534.html. [Zugriff am 21 04 2022].

[15] D. R. Moll, D. A. Schulze, M. Rusch-Rodothenous, C. Kunke und L. Scheiberl, „Wearables, Fitness-Apps und der Datenschutz," Verbraucherzentrale NRW e. V., 2017.

[16] R. Petrlic und C. Sorge, Datenschutz - Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie, Springer, 2016.

[17] S. R. Department, „statista," 076 06 2011. [Online]. Available: https://de.statista.com/statistik/daten/studie/189794/umfrage/lesen-der-datenschutzbestimmungen-im-internet/. [Zugriff am 22 04 2022].

[18] M. Al-Youssef, „Standard," 2 8 2021. [Online]. Available: https://www.derstandard.at/story/2000128639162/joe-bonusclub-soll-millionenstrafe-zahlen. [Zugriff am 22 04 2022].

[19] J. Johnson, „statista," 3 03 2021. [Online]. Available: https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/. [Zugriff am 22 04 2022].

[20] ISO, „ISO/IEC 29100," ISO, 2011.

[21] J. Frankenfield, „investopedia," 13 02 2022. [Online]. Available: https://www.investopedia.com/terms/t/tor.asp. [Zugriff am 22 04 2022].

[22] D. Cox, „BBC," 11 11 2020. [Online]. Available: https://www.bbc.com/worklife/article/20201110-the-rise-of-employee-health-tracking. [Zugriff am 21 04 2022].

[23] U. D. o. Justice, „justice.gov," 05 02 2021. [Online]. Available: https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/introduction#LegHistory. [Zugriff am 24 04 2022].

[24] EU, „edps europa," [Online]. Available: https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_de. [Zugriff am 23 04 2022].

[25] derstandard, „Standard," 14 5 2018. [Online]. Available: https://www.derstandard.at/story/2000079393003/die-sieben-wichtigsten-punkte-der-neuen-datenschutzregeln. [Zugriff am 27 04 2022].

[26] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/kapitel-1/. [Zugriff am 16 05 2022].

[27] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/artikel-3/. [Zugriff am 16 05 2022].

[28] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/artikel-4/. [Zugriff am 16 05 2022].

[29] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/artikel-5/. [Zugriff am 16 05 2022].

[30] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/artikel-6/. [Zugriff am 16 05 2022].

[31] „datenschutzexperte," [Online]. Available: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/artikel-7/. [Zugriff am 17 05 2022].

[32] wko, „wko," 20 04 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Pflichten-des-Verantwortl.html. [Zugriff am 03 05 2022].

[33] wko, „wko," 31 03 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Datensicherheit-und-Daten.html. [Zugriff am 03 05 2022].

[34] wko, „wko," 20 04 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Pflichten-des-Auftragsver.html. [Zugriff am 03 05 2022].

[35] wko, „wko," 31 03 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Dokumentationspflicht.html. [Zugriff am 03 05 2022].

[36] wko, „wko," 26 04 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung-Meldung-von-Datenschutzve.html. [Zugriff am 03 05 2022].

[37] wko, „wko," 05 05 2022. [Online]. Available: https://www.wko.at/service/wirtschaftsrecht-gewerberecht/eu-datenschutz-grundverordnung-datenschutz-folgenabschaetzu.html. [Zugriff am 10 05 2022].

[38] A. Meyermann und M. Porzelt, „Hinweise zur Anonymisierung von qualitativen Daten," Forschungsdatenzentrum (FDZ) Bildung am DIPF, Frankfurt am Main, 2014.

[39] mostlyAI; TaylorWessing, „Stellungnahme im Konsultationsverfahren zum Entwurf des Positionspapiers zur Anonymisierung Unter der Dsgvo unter Berücksichtigung der TK-Branche des Bundesbeauftragten für den Datenschutz und die Informationsfreiheit," 2020.

[40] „clarip," [Online]. Available: https://www.clarip.com/data-privacy/us-history/. [Zugriff am 26 04 2022].

[41] D. J. Solove, A Brief History of Information Privacy Law, George Washington University Law School, 2006.

[42] U. o. Michigan, „safecomputing.umich," University of Michigan, [Online]. Available: https://safecomputing.umich.edu/privacy/history-of-privacy-timeline. [Zugriff am 26 04 2022].

[43] Artikel-29-Datenschutzgruppe, „Stellungnahme 5/2014 zu Anonymisierungstechniken WP216," 2014.

[44] forschungsdatenzentrum, „forschungsdatenzentrum," [Online]. Available: https://www.forschungsdatenzentrum.de. [Zugriff am 19 05 2022].

[45] L. Willenborg und T. d. Waal, Elements of Statistical Disclosure Control, Springer, 2000.

[46] P. Planing, „Statistik Grundlagen," [Online]. Available: https://statistikgrundlagen.de/ebook/chapter/chapter-1-2/. [Zugriff am 25 05 2022].

[47] B. Lubarsky, RE-IDENTIFICATION OF "ANONYMIZED DATA", 2017.

[48] N. Li, T. Li und S. Venkatasubramanian, „t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," Department of Computer Science, Purdue University, 2007.

[49] L. Sweeney, „Computational Disclosure Control: A Primer on Data Privacy Protection," Massachusetts Institute of Technology, 2001.

[50] A. Bender, „Anwendbarkeit von Anonymisierungstechniken im Bereich Big Data," Karlsruhe Institute of Technology, Karlsruhe, 2015.

[51] H. K. TECHT, „kurier," 01 07 2019. [Online]. Available: https://kurier.at/wirtschaft/die-post-sammelt-und-verkauft-daten-zu-parteiaffinitaet/400370741. [Zugriff am 20 05 2022].

[52] Y. B. Park, „Anonymisierungsverfahren," Ludwig-Maximilians-Universität München Institut für Statistik, München, 2015.

[53] L. Sweeney und P. Samarati, „Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression".

[54] B. C. M. Fung, K. Wang, R. Chen und P. S. Yu, „Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, 2010.

[55] A. Machanavajjhala, J. Gehrke und D. Kifer, „l-Diversity: Privacy Beyond k-Anonymity," Department of Computer Science, Cornell University, 2006.

[56] L. Sweeney, „k-Anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.

[57] jupyter, „jupyter," [Online]. Available: https://jupyter.org/. [Zugriff am 24 05 2022].

[58] pandas, „pandas," [Online]. Available: https://pandas.pydata.org/about/index.html. [Zugriff am 24 05 2022].

[59] GISGeography, „GISGeography," 08 06 2021. [Online]. Available: https://gisgeography.com/wgs84-world-geodetic-system/. [Zugriff am 24 05 2022].