Anonymisierung und Pseudonymisierung von Daten für Projekte des maschinellen Lernens

Eine Handreichung für Unternehmen

www.bitkom.org

bitkom

Herausgeber

Bitkom
Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e. V.
Albrechtstraße 10 | 10117 Berlin
T 030 27576-0
bitkom@bitkom.org
www.bitkom.org

Verantwortliches Bitkom-Gremium

AK Artificial Intelligence

Projektleitung

Dr. Nabil Alsabah | Bitkom e. V.

Autoren

Patrick Aichroth | Fraunhofer-Institut für Digitale Medientechnologie (IDMT)
Verena Battis | Fraunhofer SIT Institut für Sichere Informationstechnologie
Dr. Andreas Dewes | 7scientists GmbH
Christoph Dibak | Google Germany GmbH
Vadym Doroshenko | Google Germany GmbH
Dr. Bernd Geiger | semafora systems GmbH
Lukas Graner | Fraunhofer SIT Institut für Sichere Informationstechnologie
Steffen Holly | Psoido GmbH
Prof. Dr. Michael Huth | XAIN AG & Imperial College London
Dr. Benedikt Kämpgen | Empolis Information Management GmbH
Dr. Markus Kaulartz | CMS Hasche Sigle Partnerschaft
Michael Mundt | Esri Deutschland GmbH
Dr. Hermann Rapp
Prof. Dr. Martin Steinebach | Fraunhofer SIT Institut für Sichere Informationstechnologie
Dr. Yurii Sushko | Google Germany GmbH
Dominic Swarat | Philips GmbH
Christian Winter | Software AG
Rebekka Weiß | Bitkom e.V.

Grafik und Layout

Katrin Krause | Bitkom e.V.

Titelbild

© nito | stock.adobe.com

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

1 Einleitung

Rebekka Weiß & Nabil Alsabah

Die künstliche Intelligenz ist eine junge Disziplin. Doch mit 64 ist sie doch nicht so jung, wie manche vermuten würden. Viele haben vor dem aktuellen Hype deswegen von KI nicht gehört, weil sie ihre ersten Jahrzehnte überwiegend in Forschungslaboren verbracht hat. Da hat die KI zwar mehrere Familien von Algorithmen hervorgebracht – z. B. Suche, Logik und Wissensrepräsentation. Doch von wenigen Ausnahmen abgesehen, ermöglichten diese Algorithmen keine bahnbrechenden Anwendungen in der Praxis.

Zu ihrem großen Durchbruch hat der KI jene Algorithmenfamilie verholfen, die bis dahin von der Mehrzahl der KI-Experten stiefmütterlich behandelt wurde: Das maschinelle Lernen. ML ist stellte das Paradigma der KI auf den Kopf. Nicht Regeln, sondern Daten sollen das Verhalten der KI diktieren. Will man beispielsweise einen Lernalgorithmus einsetzen, um Wölfe und Huskys in Bildern zu erkennen, so bräuchte man die Unterscheidungsmerkmale von Wölfen und Huskys nicht in Regeln zu erfassen. Vielmehr analysiert der Lernalgorithmus eine Menge von Beispielbildern beider Hundefamilien. Der Algorithmus entwickelt im Laufe der Lernphase ein generalisiertes Modell. Mit diesem Modell kann eine App später neue, bis dato nicht gesehene Bilder von Wölfen und Huskys richtig klassifizieren.

Das maschinelle Lernen hat sich in vielen Bereichen bewährt: Von der Bilderkennung in der Radiologie über Spracherkennung bei Sprachassistenten bis zur vorausschauenden Wartung in der Industrie. Dennoch: Der Grundgedanke des maschinellen Lernens ist nicht neu. Er geht vielmehr auf die fünfziger Jahre zurück. Der Siegeszug des ML liegt in der zunehmenden Verfügbarkeit von Daten *und* der rasant gestiegenen Rechenleistung begründet.

Daten sind also das Herzstück des maschinellen Lernens. Wenn wir von Daten sprechen, müssen wir die rechtlichen Rahmenbedingungen für ihre Nutzung betrachten. Insbesondere müssen wir klären, inwiefern Daten, die wir für das Trainieren von ML-Modellen nutzen, nicht nur für die Algorithmen, sondern auch aus datenschutzrechtlichen Gesichtspunkten relevant sind.

Unser Ausgangspunkt ist simpel: Aus rechtlicher Sicht muss man Daten besonders schützen und ihre Verarbeitung stark reglementieren, wenn sie personenbezogen sind. Jegliche Nutzung personenbezogener Daten unterliegt (neben weiteren rechtlichen Bestimmungen) der Datenschutzgrundverordnung. Entfernt man den Personenbezug aus den Daten, ist man also – aus rechtlicher Sicht – freier in den Nutzungsmöglichkeiten. Auch die Verschleierung des Personenbezugs dient datenschutzrechtlichen Erwägungen: Sie erhöht den Schutz der Daten, ohne aber den Anwender aus dem Korsett des Datenschutzrechts zu entlassen. Es stehen Ihnen als Entwickler zwei wichtige Werkzeuge zur Verfügung, um den Personenbezug zu verschleiern bzw. ganz zu entfernen: Sie können die Daten pseudonymisieren oder anonymisieren.

Die *Pseudonymisierung* schützt Daten, indem sie die Werte von direkten Identifikatoren (z. B. Name oder Ausweisnummer) durch Pseudonyme ersetzt. Diese Pseudonyme werden über ein geeignetes Verfahren aus dem ursprünglichen Wert generiert oder gar neu vergeben. Ein Pseudonym kann das gleiche Format wie der ursprüngliche Datentyp besitzen – z. B. ein Name

wird durch einen Künstlernamen ersetzt. Ein Pseudonym kann aber auch in einem neuen Format vorliegen – z. B. die Ausweisnummer wird durch eine zufällige Zeichenfolge ersetzt. Dabei ist es wichtig, dass die Zuordnung eindeutig ist: Für zwei identische Eingabewerte muss das gleiche Pseudonym erzeugt werden. Manche Anwendungen sind auf eine umkehrbare Pseudonymisierung angewiesen. Eine Pseudonymisierung ist dann umkehrbar, wenn man aus dem Pseudonym – auch wenn mithilfe eines zusätzlichen Schlüssels – den ursprünglichen Datenwert ableiten kann.

Pseudonymisierung wird vorwiegend eingesetzt, um sensitive Daten bei der Verarbeitung vor neugierigen Blicken zu schützen. Die Pseudonymisierung macht es lediglich schwerer, Rückschlüsse auf den ursprünglichen Datenwert zu ziehen. Da pseudonymisierte Daten eine Re-Identifikation der betroffenen Person nicht ausschließen, unterliegen sie der DS-GVO.[1]

Will man die Ableitung des ursprünglichen Datenwerts aber technisch unmöglich machen, so müsste man auf die *Anonymisierung* zurückgreifen. Anonymisierte Daten können – technisch gesehen – nicht auf individualisierte Personen zurückgeführt werden. Sie entfallen deshalb nicht dem Datenschutzrecht. Man spricht von anonymen Daten, wenn die Identifizierbarkeit eines Individuums unter Berücksichtigung sämtlicher zur Verfügung stehenden Mittel »einen unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskräften erfordern würde, sodass das Risiko einer Identifizierung de facto vernachlässigbar erschiene«.[2]

In diesem Leitfaden präsentieren wir praktische Methoden und konkrete Beispiele für die Anonymisierung und Pseudonymisierung von Daten. Der Leitfaden richtet sich insbesondere an Entwickler, die mit Fragen der Anonymisierung und Pseudonymisierung kämpfen. ↗**Kapitel 2** fasst die wichtigsten technischen Methoden und Verfahren zur A&P von Daten zusammen. ↗**Kapitel 3** geht auf die Problematik der Speicherung von Geo-Bewegungsprofilen ein. ↗**Kapitel 4** stellt die Abarbeitung von Mobilitätsdaten vor, die Google im Kontext von COVID-19 erhoben hat. ↗**Kapitel 5** diskutiert Anwendungsbeispiele für das Prinzip der *entkoppelten Pseudonyme* – ein Prinzip, welches die Rückführung von Pseudonymen erschwert.

↗**Kapitel 6** erörtert das Konzept des *föderierten Lernens* und erklärt, wie effektives maschinelles Lernen auch lokal stattfinden kann. ↗**Kapitel 7** beschreibt die Datenschutzrisiken bei Medieninhalten und empfiehlt Gegenmaßnahmen. ↗**Kapitel 8** geht auf die Anonymisierung und Pseudonymisierung medizinischer Textdaten ein. Und ↗**Kapitel 9** präsentiert das Konzept der semantischen Anonymisierung.

Wir hoffen, dass dieser Leitfaden Ihnen bei der Wahl geeigneter Verfahren helfen kann!

1 Die Pseudonymisierung ist aus datenschutzrechtlicher Sicht trotzdem ein sehr wichtiges Verfahren. Die DS-GVO erwähnt Pseudonymisierung in den Artikeln 25, 26 und 40 sowie in Erwägungsgrund 28 und hebt die Wichtigkeit dieses Verfahrens bei der Verarbeitung personenbezogener Daten hervor.

2 Vgl. EuGH, a.a.O., Rn. 46.

2 Technische Werkzeuge für die Anonymisierung und Pseudonymisierung von Daten

2 Technische Werkzeuge für die Anonymisierung und Pseudonymisierung von Daten

Andreas Dewes, Martin Steinebach, Patrick Aichroth, Christian Winter, Benedikt Kämpgen

Anonymisierung durch Aggregation, Hinzufügen von Rauschen und Synthese, Angriffe auf anonyme Daten, entkoppelte Pseudonyme, Anonymisierung von Multimedia Daten, On-Prem-Analyse und Dezentralisierung

2.1 Anonymisierung strukturierter Daten

Strukturierte Datensätze bestehen aus einzelnen Datenpunkten. In tabellarischer Darstellung eines Datensatzes entspricht ein Datenpunkt einer Tabellenzeile. Jeder Datenpunkt des Datensatzes enthält Attribute, die konkrete Werte besitzen.

Es gibt eine Reihe von Verfahren, mit denen strukturierte Daten anonymisiert werden können. Welches Verfahren anwendbar ist, hängt u.a. von der Art der zu anonymisierenden Daten, dem geplanten Verwendungszweck der Daten sowie den technischen und organisatorischen Rahmenbedingungen der Datennutzung ab. In den folgenden Abschnitten erläutern wir folgende Ansätze:

- **Aggregationsbasierte Verfahren**: Hierbei werden einzelne Datenpunkte des Ursprungs-Datensatzes zu Gruppen aggregiert, wodurch eine Re-Identifikation sowie die Bestimmung bzw. zuverlässige Schätzung von Attributwerten einzelner Personen erschwert wird.

- **Zufallsbasierte Verfahren**: Hierbei werden einzelne Attribute zufallsbasiert so verändert, dass eine Re-Identifikation und die zuverlässige Schätzung von Attributwerten einzelner Personen erschwert wird.

- **Synthesebasierte Verfahren**: Hierbei wird zunächst ein statistisches Modell der Ursprungsdaten gebildet. Anhand dieses Modells werden anschließend neue, synthetische Daten generiert, welche die Ursprungsdaten möglichst gut nachbilden aber keinen Personenbezug mehr aufweisen sollen.

Die Verfahren nutzen unterschiedliche Ansätze, um die Anonymität der transformierten Daten im Rahmen eines Risikomodells nachzuweisen. Bei jedem Verfahren ist es daher wichtig, dieses Risikomodell und seine Grenzen zu kennen, um die Eignung des Verfahrens für einen gegebenen Anwendungsfall bewerten zu können. Oft ist es zudem möglich und sinnvoll, mehrere dieser Verfahren zu kombinieren, um eine stärkere Anonymität zu erreichen. Die Verfahren können

zudem zu unterschiedlichen Zeitpunkten auf Daten angewandt werden. Man unterscheidet in der Praxis grob drei Szenarien:

- **Statische Anonymisierung**: Hierbei wird ein bestehender, unveränderlicher und vollständig bekannter Datensatz nach vorher festgelegten Kriterien anonymisiert.
  Beispiel: Eine Tabelle mit Patientendaten wird durch Aggregation einmalig anonymisiert; die anonymisierten Daten werden zu Forschungszwecken benutzt.

- **Dynamische Anonymisierung**: Hierbei wird ein kontinuierlicher Strom von Daten nach vorher festgelegten Kriterien anonymisiert.
  Beispiel: Ein Strom aus Positionsdaten wird rauschbasiert anonymisiert; die anonymisierten Daten werden in Echtzeit weiterverarbeitet.

- **Interaktive Anonymisierung**: Hierbei wird ein (meist) statischer Datensatz nach dynamisch festgelegten Kriterien interaktiv anonymisiert.
  Beispiel: Das Ergebnis einer durch einen Anwender definierten SQL-Anfrage auf eine Datenbanktabelle wird rauschbasiert anonymisiert bevor es zurückgegeben wird.

Anonymisierung durch Aggregation

Aggregationsbasierte Verfahren gruppieren einzelne Datenpunkte des Ursprungsdatensatzes. Die Gruppierung erfolgt hierbei so, dass die Nutzbarkeit der Daten möglichst erhalten bleibt, aber das Risiko der Re-Identifikation und der Bestimmung von Attributwerten einzelner Personen reduziert wird. Aggregationsbasierte Anonymisierung wird seit langem angewandt und u.a. vom statischen Bundesamt eingesetzt.

Üblicherweise werden bei diesen Verfahren identifizierende Merkmale entweder generalisiert oder mittels sogenannter Mikroaggregation innerhalb der Gruppen durch repräsentative Werte ersetzt. Bei der Generalisierung wird beispielsweise das genaue Alter durch Fünfjahresintervalle ersetzt oder der genaue Beruf durch eine Qualifikationsstufe. Hier richtet die Gruppierung sich nach den vergröberten Merkmalen. Bei der Mikroaggregation hingegen werden grundsätzlich zuerst die Gruppen festgelegt und danach wird beispielsweise das individuelle Alter durch den Median des Alters innerhalb der Gruppe ersetzt.

k-Anonymität

Ende der 90er Jahre wurde mit *k*-Anonymität ein formelles Kriterium zur Bewertung der Anonymität aggregierter Daten eingeführt. Hierbei werden die Attribute eines Datensatzes zunächst unterteilt in nicht-sensible und sensible Attribute. Die sensiblen Attribute wie etwa Krankheiten stellen besonders schützenswerte Informationen zu Personen dar. Die nicht-sensiblen Attribute sind allgemeine Personenmerkmale wie das Alter und das Geschlecht. Die nicht-sensiblen Attribute werden hierbei oft als Quasi-Identifikatoren bezeichnet, da sie in Kombination innerhalb eines Datensatzes eindeutig sein und gleichzeitig leicht mit anderen Datensätzen verknüpft werden können und damit zur Re-Identifikation einzelner Personen genutzt werden

könnten. Anschließend wird der Datensatz nach allen nicht-sensiblen Attributen gruppiert, wobei die Werte der sensiblen Attribute von den zugehörigen Datenpunkten losgelöst und der Gruppe als Ganzes zugeordnet werden. Der resultierende Datensatz wird als *k*-anonym bezeichnet, wenn in jeder so gebildeten Gruppe mindestens *k* einzelne Datenpunkte vorhanden sind. Die Anonymität der Daten soll dadurch erreicht werden, dass keine eindeutige Zuordnung zwischen sensiblen Attributwerten und einzelnen Datenpunkten von Personen in der Gruppe möglich ist.

Nach der Veröffentlichung des Konzeptes der *k*-Anonymität wurden recht schnell Schwächen in dem Schema entdeckt. Beispielsweise ist es denkbar, dass alle einer Gruppe zugehörigen Datenpunkte den gleichen Wert eines sensiblen Attributs aufweisen. Eine Gruppierung schützt damit die Daten der Personen in der Gruppe nicht vor Aufdeckung ihres Attributwerts, da dieser für alle Mitglieder der Gruppe identisch ist und ein Angreifer mit Sicherheit auf diesen Wert schließen kann, selbst ohne die genaue Zuordnung der in diesem Fall ununterscheidbaren Einzelwerte auf Gruppenmitglieder zu kennen. Dieses Problem kann behoben werden, indem zusätzlich zur *k*-Anonymität ein weiteres Kriterium eingeführt wird, welches wir im nächsten Abschnitt diskutieren.

_l_-Diversität

Zur Behebung des vorher beschriebenen Problems kann *k*-Anonymität durch den Begriff der *l*-Diversität erweitert werden: Hierbei wird für jede gebildete Gruppe die Anzahl der unterschiedlichen Attributwerte erfasst. Wenn alle Gruppen mindestens *l* verschiedene Attributwerte beinhalten, bezeichnet man den Datensatz als *l*-divers. *l*-Diversität schützt Personen vor Offenlegung ihrer sensiblen Attributwerte, indem ausgeschlossen wird, dass einzelne Gruppen weniger als *l* verschiedene sensible Attributwerte beinhalten. Dies ermöglicht Mitgliedern der Gruppe, glaubhaft abzustreiten, dass sie einen gegebenen Attributwert besitzen.

l-Diversität kann als alleiniges Kriterium unabhängig von *k*-Anonymität genutzt werden. Ein *l*-diverser Datensatz ist immer auch mindestens *l*-anonym, da für l-Diversität mindestens *l* Einträge in einer gegebenen Gruppe notwendig sind. Bestimmte Attributtypen wie z. B. numerische Attribute müssen für den Einsatz von *l*-Diversität zusätzlich durch Gruppierung quantisiert werden; alternativ muss das Diversitätskriterium für diese Attribute angepasst werden. *l*-Diverse Datensätze sind weiterhin angreifbar, da bei großen Gruppengrößen weiterhin eine starke Konzentration auf einzelne Attributwerte möglich ist: So ist z. B. eine Gruppe mit 1000 Mitgliedern auch 2-divers, wenn 999 Mitglieder den gleichen Attributwert haben und lediglich ein Mitglied einen abweichenden Wert besitzt. In diesem Fall kann ein Angreifer mit großer Erfolgswahrscheinlichkeit den Attributwert einer Person aus der Gruppe durch Raten richtig treffen; einzelne Personen haben damit nur eine sehr geringe plausible Abstreitbarkeit. Dieses Problem kann wiederum behoben werden, indem ein weiteres Kriterium hinzugezogen wird.

t-Ähnlichkeit

Zur Behebung des beschriebenen Konzentrationsproblems kann das Kriterium der *t*-Ähnlichkeit (englisch »*t*-closeness«) für einzelne Gruppen im aggregierten Datensatz eingeführt werden. Dieses Kriterium erfasst für jede Gruppe, wie stark die Verteilung der sensiblen Attributwerte über diese Gruppe von der Verteilung über den gesamten Datensatz abweicht. Der Grad der Abweichung ist hierbei keine eindeutig definierte Größe, es werden zur Messung vielmehr häufig Metriken wie die Kullback-Leibler Divergenz oder die sogenannte »earth mover's distance« herangezogen. Ein Datensatz ist *t*-ähnlich, wenn der Wert dieser Metrik in jeder Gruppe maximal *t* beträgt. Durch die Beschränkung der Abweichung zwischen bedingter und unbedingter Verteilung der sensiblen Attributwerte, wird die Anonymität einzelner Personen in der Gruppe besser geschützt als bei der Nutzung von *l*-Diversität. Jedoch bestehen auch hier weiterhin Risiken, da es oft schwierig ist einen adäquaten Wert für *t* zu definieren und das Re-Identifikations- und Attributbestimmungsrisiko einzelner Gruppen je nach der Verteilung der Attributwerte stark unterschiedlich ausfallen kann.

Implementierung

Die Erstellung eines *k*-anonymen, *l*-diversen oder gar *t*-ähnlichen Datensatzes ist nicht immer einfach, insbesondere wenn die zugrunde liegenden Ursprungsdaten eine hohe Anzahl an sensiblen und nicht-sensiblen Attributen beinhalten. In der Praxis wurden eine Reihe von Verfahren entwickelt, um iterativ solche Datensätze zu generieren. Da eine Gruppierung einzelner Datenpunkte beliebig erfolgen kann, steigt die Anzahl der möglichen Gruppierungen hierbei exponentiell mit der Anzahl der Datenpunkte an. Oft unterliegt die Gruppierung zusätzlich einem Optimierungskriterium; beispielsweise ist es vielfach wünschenswert, einzelne Gruppen aus sehr ähnlichen Datenpunkten zu bilden, da dies oft die statistische Auswertung vereinfacht.

Einer der beliebteren Ansätze hierfür ist der sogenannte »Mondrian-Algorithmus«. Dieser gruppiert alle Datenpunkte zunächst in eine einzelne Gruppe. Diese wird anschließend unter Berücksichtigung des gewählten Anonymitätskriteriums (*k*-Anonymität, *l*-Diversität oder *t*-Ähnlichkeit) in zwei neue Gruppen aufgeteilt. Für jede so entstehende Gruppe wird der Prozess der Teilung solange wiederholt, bis die neu entstehenden Gruppen das Anonymitätskriterium erfüllen.

Ein weiterer etablierter Ansatz ist die MDAV-Methode (MDAV steht für »Maximum Distance to Average Vector«), welche in die Domäne der Mikroaggregierung gehört und daher insbesondere für numerische Attribute geeignet ist. Hier werden die Datenpunkte gemäß ihres Abstands zueinander gruppiert, wobei die Gruppen nach Möglichkeit nicht mehr als *k* Elemente enthalten sollen. Dazu werden zuerst solche Gruppen gebildet, die möglichst weit von der »Mitte« entfernt sind, damit am Ende keine Datenpunkte am »Rand« übrigbleiben, die nicht sinnvoll gruppiert werden können. Schrittweise werden weitere Datenpunkte gruppiert, bis auch die letzten Datenpunkte in der Mitte eine Gruppe bilden.

Anwendbarkeit aggregationsbasierter Anonymisierung

Aggregationsbasierte Verfahren sind prinzipiell einfach zu implementieren und können auf statische sowie (mit Einschränkungen) auf dynamische Datensätze angewandt werden.

Vorteile aggregationsbasierter Anonymisierung

Aggregationsbasierte Verfahren haben den Vorteil, dass sie oft einfach strukturiert und gut verständlich sind, was eine Überprüfung vereinfacht. Die entstehenden aggregierten Daten können zudem einfach interpretiert werden und beinhalten im Gegensatz zu rauschbasiert anonymisierten oder synthetischen Daten keine zufälligen Veränderungen einzelner Attributwerte oder Attributwertkombinationen, was eine Analyse und Interpretation stark vereinfachen kann.

Nachteile aggregationsbasierter Anonymisierung

Die Bildung geeigneter Gruppen für die Aggregierung ist ein mathematisch hochkomplexes Problem, das im Regelfall nicht exakt gelöst werden kann. Dementsprechend werden oft heuristische Verfahren eingesetzt, die mithilfe eines geeigneten Optimierungsverfahrens eine Gruppierung der Daten durchführen. Gerade bei Datensätzen mit sehr vielen Attributwerten oder Datenpunkten kann diese Gruppierung sehr komplex sein und die Anwendbarkeit der aggregationsbasierten Anonymisierung beschränken. Eine dynamische oder interaktive Anonymisierung ist mithilfe von aggregationsbasierten Verfahren zudem nur unter Einschränkungen möglich. Im Rahmen der interaktiven Anonymisierung kann eine mehrfache Aggregation von Daten zu einem Verlust der Anonymität führen. Bei der dynamischen Anonymisierung kann hingegen eine Gruppierung nicht unter Berücksichtigung des Gesamtdatensatzes erfolgen; der Informationsverlust ist in diesem Fall daher oft höher als bei der Anonymisierung eines statischen Datensatzes.

Anonymisierung durch Hinzufügen von Rauschen

Bei der rauschbasierten Anonymisierung werden Attributwerte eines Datensatzes durch künstlich erzeugtes, statistisches Rauschen zufällig verändert. Dies bewirkt, dass der wirkliche Wert eines gegebenen Attributs nicht mehr mit Sicherheit bestimmt werden kann, was eine plausible Abstreitbarkeit und Anonymität für die betroffene Person schafft. Wie bei anderen Anonymisierungsverfahren reduziert sich auch hier die Nutzbarkeit der Daten, da der Datensatz verfälscht wird. Analysen des Datensatzes müssen die Veränderung als Störeffekt berücksichtigen, dies kann je nach beabsichtigter Nutzung problematisch sein. Je nach Verfahren können bei der Veränderung von Attributwerten auch ungültige Daten erzeugt werden, die außerhalb von plausiblen oder erlaubten Attributwertkombinationen liegen. Dies kann zum einen die Analyse erschweren und zum anderen die Anonymität der Daten schwächen, da so eventuell weitergehende Rückschlüsse jenseits der Anonymitätsgarantien des Rauschmodells auf die Ursprungsdaten möglich werden.

Es gibt eine Vielzahl von Verfahren, die zur rauschbasierten Veränderung von Daten eingesetzt werden können. Welches Verfahren anwendbar ist, hängt maßgeblich von der Art der zu anonymisierenden Daten und der beabsichtigten Nutzung ab. Der Nachweis der Anonymität kann über statistische Verfahren erfolgen. Insbesondere moderne Bewertungsansätze wie Differential Privacy können als Maßstab zur Beurteilung des Effekts der rauschbasierten Veränderung der Daten herangezogen werden.

Anwendbarkeit rauschbasierter Anonymisierung

Rauschbasierte Verfahren können für die interaktive, statische als auch dynamische Anonymisierung von Daten eingesetzt werden. Für die interaktive Anonymisierung von Daten durch Hinzufügen von Rauschen existieren mehrere kommerzielle wie nichtkommerzielle Lösungen, ebenso für die Anonymisierung von dynamischen Daten.

Vorteile rauschbasierter Anonymisierung

Rauschbasierte Verfahren erfreuen sich zunehmender Beliebtheit da sie üblicherweise mit modernen Verfahren wie *Differential Privacy* analysiert werden können und damit gute Anonymitätsgarantien bieten. Sie können zudem oft auf einzelne Datenpunkte oder – im Falle der interaktiven Anonymisierung – Abfrageergebnisse angewandt werden, ohne hierfür den Gesamtdatensatz betrachten zu müssen, was eine Umsetzung gerade bei sehr großen Datenmengen erleichtert. Rauschbasierte Verfahren bewahren zusätzlich oft das Format und die Struktur des Originaldatensatzes, was eine Analyse der Daten vereinfachen kann.

Nachteile rauschbasierter Anonymisierung

Durch Hinzufügen von Rauschen werden die Ursprungsdaten verändert. Diese Veränderung muss in allen Analysen, die auf diesen verrauschten Daten basieren berücksichtigt werden. Je nach eingesetzter Methodik kann das Hinzufügen von Rauschen zu unrealistischen oder invaliden Daten führen, insbesondere wenn mehrere Attributwerte anonymisiert werden sollen. Wenn rauschbasierte Verfahren für die interaktive Anonymisierung eingesetzt werden, muss zudem durch geeignete Maßnahmen sichergestellt werden, dass ein Angreifer das Rauschen nicht durch mehrfache Abfragen und anschließendes statistisches Mitteln reduzieren kann. Robuste Verfahren, um solche Angriffsarten zu verhindern sind ein aktiver Forschungsgegenstand, in verschiedenen existierenden Ansätzen wurden in der Vergangenheit mehrfach Schwachstellen identifiziert.

Anonymisierung durch Synthese

Datensynthese anonymisiert Daten in einem zweistufigen Verfahren:

- Zunächst wird ein statistisches Synthesemodell an die Ursprungsdaten angepasst.
- Mithilfe dieses Synthesemodells werden neue, synthetische Daten generiert.

Die Anonymität der synthetischen Daten kann hierbei durch mehrere Arten sichergestellt werden. Entweder können die Ursprungsdaten oder die synthetischen Daten zusätzlich durch ein anderes Anonymisierungsverfahren geschützt werden, oder bei der Generierung des Synthesemodells können entsprechende Anonymitätsgarantien vorgesehen werden. Im letzteren Fall kann z. B. durch Hinzufügen von Rauschen oder durch Einschränkung der Lernrate des Synthesemodells verhindert werden, dass dieses zu viele Informationen von einzelnen Datenpunkten des Ursprungsdatensatzes extrahiert. Die Anonymität von Daten ist bei der Synthese schwerer nachzuweisen als bei anderen, direkten Verfahren, da hierzu die innere Funktionsweise des Synthesemodells und des zugehörigen Lernverfahrens, welches die Parameter des Modells anhand der Ursprungsdaten generiert, untersucht werden müssen.

Oberflächlich betrachtet mögen synthetische Daten sicherer erscheinen als anderweitig anonymisierte Daten, da keine direkten Beziehungen zwischen einzelnen Datenpunkten der Ursprungsdaten und den synthetischen Daten bestehen. Dieser Eindruck trügt jedoch oft, da aus der statistischen Verteilung von synthetischen Daten ebenfalls Rückschlüsse auf einzelne Personen gezogen werden können und es, in Abhängigkeit des Syntheseverfahrens sogar möglich sein kann, einzelne Personen zu re-identifizieren in dem Sinne, dass auf ihre Präsenz in den Ursprungsdaten geschlossen werden kann, sowie dass zuverlässige Schätzungen von Attributwerten der Personen möglich sind.

Anwendbarkeit von Synthese

Synthesebasierte Anonymisierung kann auf statische sowie dynamische Datensätze angewandt werden. Bei der Anwendung auf statische Datensätze wird hierbei zunächst basierend auf dem Datensatz ein Synthesemodell generiert, mit diesem werden anschließend neue Daten synthetisiert. Bei dynamischen Datensätzen wird das Synthesemodell hingegen kontinuierlich an neue Daten angepasst und auch die Synthese erfolgt kontinuierlich. Oft wird hierbei eine Einlaufphase benötigt, in der keine Daten synthetisiert werden und die genutzt wird, um anhand der eintreffenden dynamischen Daten ein erstes Synthesemodell bilden zu können.

Vorteile von Synthese

Synthese kann Datensätze erzeugen, welche die Struktur und das Format der ursprünglichen Daten gut widerspiegeln. Da keine direkte Beziehung zwischen einzelnen Datenpunkten der Ursprungsdaten und der synthetischen Daten besteht, ist es für einen Angreifer zunächst schwieriger, durch direkten Vergleich Rückschlüsse auf einzelne Personen in den synthetischen Daten zu finden.

Nachteile von Synthese

Um die Anonymität der Daten zu bewahren, müssen das Synthesemodell oder die synthetischen Daten mit einem geeigneten Anonymitätskriterium beschränkt werden. Werden hierfür moderne Verfahren wie Differential Privacy genutzt, reduziert sich die Genauigkeit der synthetisierten Daten linear mit der Anzahl der Parameter des Synthesemodells. Je mehr Attribute ein Daten-

satz besitzt, umso schwieriger ist es, einen realistischen synthetischen Datensatz zu generieren, der gleichzeitig gute Anonymitätsgarantien bietet. So hat ein Datensatz mit 16 binären (ja/nein) Attributwerten bereits 2^{16}=65.536 Attributkombinationen, die theoretisch für ein realistisches Synthesemodell hinsichtlich ihrer Wahrscheinlichkeitsverteilung analysiert werden müssen. Da die Zahl der hierfür nötigen Parameter mit der Anzahl der Attribute im Datensatz exponentiell wächst, können Syntheseverfahren meist nur einen kleinen Ausschnitt der Wahrscheinlichkeitsverteilung dieser Ursprungsdaten erfassen und modellieren. Dies führt dann dazu, dass synthetische Daten zwar auf der Ebene einzelner Attributwerte den Ursprungsdaten ähneln, komplexe Attributbeziehungen jedoch verlorengehen. Wenn das Synthesemodell nicht explizit steuerbar ist, kann zudem oft nicht nachvollzogen werden, welche Eigenschaften der ursprünglichen Daten in den synthetischen Daten erhalten werden oder bei der Synthese verlorengehen.

Risikobewertung und Angriffe auf anonyme Daten

Anonyme Datensätze können auf unterschiedliche Arten angegriffen werden, um sie zu de-anonymisieren. Ein Angreifer verfolgt hierbei üblicherweise eines oder mehrere der folgenden Ziele:

- Herauszufinden, ob die Daten einer spezifischen Person Teil des Ursprungsdatensatzes waren, aus dem die anonymen Daten generiert wurden.
- Herauszufinden, welche Datenpunkte im anonymisierten Datensatz die Daten einer spezifischen Person beinhalten.
- Vorhersagen über die Werte von Attributen einer spezifischen Person zu machen.

Welche dieser Angriffsszenarien relevant sind und ein Risiko darstellen, hängt vom Einzelfall ab. In manchen Fällen kann bereits die Information, dass die Daten einer gegebenen Person im Ursprungsdatensatz vorhanden waren, schädlich für diese Person sein. Es ist zudem für einen Angreifer nicht immer nötig, eine sichere Vorhersage einzelner Attributwerte von Personen zu machen, da es bereits ausreichen kann, Attributwerte mit einer gewissen Wahrscheinlichkeit vorhersagen zu können.

Angriffsverfahren, mit denen diese Ziele erreicht werden können, müssen im Normalfall auf die jeweils verwendeten Anonymisierungsverfahren angepasst werden: So müssen per Datensynthese anonymisierte Datensätze anders angegriffen werden als solche, die mithilfe von Aggregation oder Hinzufügen von Rauschen generiert wurden. Weiterhin benötigt der Angreifer üblicherweise Kontextinformationen zu einzelnen Personen, die er mit den anonymisierten Daten abgleichen kann. Prinzipiell gilt, dass es für einen Angreifer umso einfacher ist, Personen in einem anonymisierten Datensatz zu de-anonymisieren, je mehr relevante Kontextinformationen dieser über die Personen hat, und je genauer er das zur Anonymisierung eingesetzte Verfahren kennt. Die Geheimhaltung des Anonymisierungsverfahrens kann daher eine sinnvolle Sicherheitsmaßnahme sein; die Kenntnis des Verfahrens allein sollte die Wahrscheinlichkeit der erfolgreichen De-Anonymisierung für einen Angreifer jedoch nicht wesentlich erhöhen.

Um eine quantitative Aussage zur Wahrscheinlichkeit der Re-Identifikation einer Person sowie der Vorhersage von Attributwerten der Person zu machen, wird oft ein formelles Angriffsmodell definiert und mit einem Testdatensatz evaluiert. Ein solches Modell kann für einen gegebenen Datensatz z. B. folgende Metriken evaluieren:

- Die Genauigkeit, mit der ein Angreifer eine spezifische Person in den anonymisierten Datensatz re-identifizieren kann. Re-Identifikation schließt hierbei auch ein, dass ein Angreifer vorhersagen kann, ob die Daten einer spezifischen Person in dem Ursprungs-Datensatz, auf dem der anonymisierte Datensatz basiert, vorhanden waren. Diese Definition erfasst damit auch indirekte Anonymisierungsverfahren wie die Datensynthese, bei der es oft nicht möglich ist, einen einzelnen Datenpunkt auf eine Person zu beziehen, aber gegebenenfalls trotzdem ein Bezug einzelner Elemente des Datensatzes zu der Person hergestellt werden kann.

- Die Genauigkeit, mit der ein Attributwert einer spezifischen Person basierend auf der Kenntnis des anonymen Datensatzes sowie zusätzlicher Attribute der Person vorhergesagt werden kann, gegebenenfalls in Relation zu der Genauigkeit der Vorhersage ohne Kenntnis des Datensatzes oder bei Kenntnis einer Version eines anonymisierten Datensatzes, der die Daten der betreffenden Person nicht beinhaltet.

Bei der Risikobewertung von anonymen Datensätzen werden häufig mehrere Szenarien betrachtet. Hierbei werden Angreifer simuliert, die über unterschiedlich detaillierte Kenntnisse des Anonymisierungsverfahrens verfügen, unterschiedlich detaillierte Kontextinformationen sowie unterschiedliche Ressourcen zur Verfügung haben. So kann die Anonymität eines Datensatzes gegenüber verschieden mächtigen Angreifern untersucht werden. Die identifizierten Risiken können dann entweder durch Anpassung des Anonymisierungsverfahrens selbst, oder durch zusätzliche technisch-organisatorische Schutzmaßnahmen reduziert werden, um eine rechtskonforme Verarbeitung zu gewährleisten.

Die Szenarien sowie Angriffsverfahren, die betrachtet werden, müssen üblicherweise dem zugrunde liegenden Anonymisierungsverfahren sowie dem betrachteten Datensatz angepasst werden. In den folgenden Abschnitten beschreiben wir daher für jedes der vorher diskutierten Anonymisierungsverfahren relevante Angriffsverfahren und Risikoszenarien, und liefern Ansätze für die Untersuchung und Bewertung relevanter Risiken.

Angriffe auf aggregierte Daten

Auf aggregierte Daten sind je nach den verfügbaren Kontextinformationen verschiedene Angriffsszenarien denkbar. So können über die in den Abschnitten zu aggregationsbasierten Verfahren diskutierten Ansätze genutzt werden, um basierend auf bekannten Daten einer Person Rückschlüsse auf sensible Attributwerte dieser Person zu ziehen. Sobald die Verteilung eines sensiblen Attributs in einer Gruppe von der Verteilung des Attributs in dem Gesamtdatensatz abweicht, kann ein Angreifer eine statistische Vorhersagen über den Attributwert einer Person erlangen, wenn er weiß, dass die Person im Datensatz enthalten ist und einer gegebenen Gruppe angehört. Ob dies eine Verletzung der Anonymität der betroffenen Person bedeutet,

hängt hierbei vom Ausmaß des Informationsgewinns des Angreifers ab: Konnte dieser z. B. den Attributwert der Person vor Kenntnis der anonymen Daten mit einer Genauigkeit von 30 % vorhersagen, nach Kenntnisnahme der Daten jedoch mit einer Genauigkeit von 95 %, so kann in vielen Fällen von einer potenziell schädlichen Auswirkung auf die Privatsphäre der Person ausgegangen werden. In anderen Fällen kann bereits eine leichte Erhöhung der Vorhersagegenauigkeit schädlich sein, die Festlegung eines konkreten quantitativen Kriteriums kann daher nur im Einzelfall erfolgen.

Angriffe auf mit Rauschen anonymisierte Daten

Rauschbasiert anonymisierte Daten können auf verschiedene Arten angegriffen werden. Generiert das Rauschverfahren beispielsweise unrealistische oder sehr unwahrscheinliche Attributwerte kann dies einem Angreifer erlauben, genauere Vorhersagen zum möglichen Wert eines Attributs zu machen als durch das Rauschmodell vorgesehen ist. Ist es einem Angreifer zudem möglich, wiederholt einzelne Datenpunkte durch das Rauschverfahren zu anonymisieren, beispielsweise im Rahmen einer dynamischen oder interaktiven Anonymisierung, kann er zudem das Rauschen durch statistisches Mitteln reduzieren und so ebenfalls die Anonymitätsgarantie des Rauschmodells umgehen. Auf ähnliche Weise kann ein Angreifer korrelierte Attributwerte nutzen, um durch geschicktes statistisches Mitteln das effektive Rauschen eines Attributwerts zu reduzieren. Schließlich sind auch Angriffe auf das Rauschverfahren selbst möglich, beispielsweise wenn deterministische, pseudozufällige Verfahren benutzt werden, um das Rauschen für einzelne Datenpunkte zu erzeugen und der Angreifer das hierfür zugrunde liegende Verfahren nachbilden kann.

Angriffe auf synthetische Daten

Synthetische Daten scheinen zunächst schwerer angreifbar zu sein als Daten, die mit rauschbasierten Verfahren oder durch Aggregation anonymisiert wurden. Jedoch können auch solche Daten Rückschlüsse über einzelne Personen erlauben. Beispielsweise nutzen manche Syntheseverfahren direkt Attributwerte des Originaldatensatzes zur Generierung der synthetischen Daten. Handelt es sich hierbei um relativ eindeutige Werte (z. B. numerische Werte mit hoher Genauigkeit und Spezifität) kann die Präsenz eines Wertes in den synthetischen Daten einem Angreifer erlauben, auf die Präsenz einer spezifischen Person in den Ursprungsdaten zu schließen. Die statistische Analyse der synthetischen Daten kann einem Angreifer zudem ähnlich wie bei Aggregationsverfahren erlauben, Vorhersagen zu Attributwerten einzelner Personen zu treffen.

Fazit

Es existiert eine Vielzahl von Verfahren, mit denen Daten in der Praxis anonymisiert werden können. Welcher Ansatz für einen gegebenen Fall anwendbar ist, hängt maßgeblich von dem Format der zu schützenden Daten und der beabsichtigten Nutzung ab. Je nach Anwendung können statische, dynamische oder interaktive Verfahren genutzt werden. Neben der technischen Eignung des Anonymisierungsverfahrens sollte immer untersucht werden, ob das Verfah-

ren geeignet ist, alle bekannten und relevanten Risiken für Personen, deren Daten anonymisiert werden sollen, effektiv zu reduzieren. Verfahren, die auditierbar sind und über statistische Sicherheitsbeweise verfügen, sollten gegenüber intuitiven oder heuristischen Ansätzen bevorzugt werden. Alle gewählten Parameter des Anonymisierungsverfahrens sollten anhand nachvollziehbarer, relevanter Kriterien gewählt werden. Die anonymisierten Daten sollten zusätzlich von unabhängiger Stelle auf mögliche Risiken geprüft werden.

2.2 Pseudonymisierung

Pseudonymisierung ist eine weitere Möglichkeit, sensible und personenbezogene Daten bei der Verarbeitung zu schützen. Im Gegensatz zur Anonymisierung bleibt hierbei jedoch der Personenbezug mittelbar erhalten und die pseudonymisierten Daten unterliegen weiterhin der DSGVO. Dort wird Pseudonymisierung als technisch-organisatorische Maßnahme (TOM) betrachtet, die ähnlich zur Verschlüsselung von Daten das Risiko für Betroffene bei der Verarbeitung personenbezogener Daten senkt. Im Gegensatz zu verschlüsselten Daten, die zur Nutzung zunächst entschlüsselt werden müssen, können pseudonymisierte Daten ohne De-Pseudonymisierung verarbeitet werden, da meist nur direkte Identifikationsmerkmale entfernt werden, andere Bestandteile der Daten jedoch erhalten bleiben. Pseudonymisierte Daten bieten daher oft einen akzeptablen Kompromiss zwischen dem Schutz sensibler Daten auf der einen Seite und der Erhaltung der Nutzbarkeit der Daten auf der anderen Seite. Verarbeitungsvorgänge, die mit direkt personenbezogenen Daten aus Risikosicht nicht vertretbar sind, können mit pseudonymisierten Daten eventuell durchgeführt werden.

In der Praxis existiert eine Vielzahl von Verfahren für die Pseudonymisierung. Eine Klassifikation ist beispielsweise anhand des eingesetzten Mechanismus möglich: Einige Verfahren nutzen kryptographische Techniken wie Hashing oder Verschlüsselung, um aus Ursprungsdaten Pseudonyme abzuleiten. Andere Verfahren setzen auf zufallsgenerierte oder manuell erstellte Werte, um Pseudonyme abzuleiten, welche in Tabellen gespeichert werden. Schließlich ist auch eine Kombination beider Techniken möglich.

Zusätzlich können wir Verfahren anhand ihres Abbildungsverhaltens unterscheiden: Verfahren wie die formaterhaltende Verschlüsselung erlauben die direkte Rückberechnung eines Ursprungswertes aus einem Pseudonym mithilfe einer zusätzlichen Information, z. B. eines kryptographischen Schlüssels. Andere Verfahren, wie z. B. Hashing von Eingabewerten, sind hingegen nicht umkehrbar und generieren teilweise auch keine global eindeutigen Pseudonyme (bei modernen Hashing-Verfahren ist die Wahrscheinlichkeit einer Kollision zweier Pseudonyme jedoch absolut vernachlässigbar).

Weiterhin können Verfahren anhand ihrer Anwendbarkeit auf unterschiedliche Datentypen unterschieden werden. Viele Pseudonymisierungstechniken werden auf direkte Identifikationsmerkmale wie numerische IDs oder Namen angewandt. Pseudonymisierung im weiteren Sinne kann jedoch auch auf andere Attributmerkmale eines Datensatzes angewandt werden. So können z. B. numerische Werte, Datumsangaben oder strukturierte Daten wie beispielsweise

IP-Adressen pseudonymisiert werden. Hierbei gibt es Techniken wie die formaterhaltende Verschlüsselung (welche oft als Pseudonymisierungsverfahren betrachtet wird, da es eine 1:1 Zuordnung zwischen Ursprungsdaten und pseudonymisierten Daten gibt), welche das Ursprungsformat der Daten bei der Pseudonymisierung erhalten kann. Andere Verfahren können zudem bestimmte Strukturen wie Hierarchien in den pseudonymisierten Daten erhalten. Welche spezifischen Techniken angewandt werden können, hängt vom Anwendungsfall ab: Um die Nutzbarkeit der pseudonymisierten Daten zu erhalten, müssen einerseits bestimmte Eigenschaften bei der Pseudonymisierung erhalten bleiben. Andererseits muss vermieden werden, dass Personen in den pseudonymisierten Daten einfach identifiziert werden können.

2.3 Funktionstrennung und »entkoppelte Pseudonyme«

Die Verwendung von Pseudonymen ist besonders für Systeme interessant, wo für Datenanalysen eine eindeutige, differenzierte Zuordnung zu Personen (oder anderen schützenswerten Identitäten wie Objekten, Organisationen etc.), aber keine Kenntnis der dahinterliegenden realen Identitäten erforderlich ist. Dazu gehören zum Beispiel Personalisierungs- und Empfehlungsdienste oder Systeme, bei denen Daten zum Zweck einer gemeinsamen Modellbildung ausgetauscht werden müssen.

Für solche Fälle kann die Erzeugung von Pseudonymen auch über einen Datentreuhänderdienst (Trusted Third Party) im Sinne einer Funktionstrennung realisiert werden: Der Dienst sorgt dafür, dass der Datenempfänger alleine die so »entkoppelten Pseudonyme« keiner Realidentität mehr zuordnen kann (immer vorausgesetzt, die dem Pseudonym zugeordneten Daten wurden ebenfalls so aufbereitet bzw. anonymisiert, dass sie nicht doch indirekt auf die Realidentität verweisen). Dieser Ansatz setzt allerdings ein hohes Vertrauen in die Motive und Sicherheitskompetenzen des Datentreuhänderdienstes voraus: Kommen die entsprechenden Daten dort z. B. durch das Fehlverhalten von Mitarbeitern oder durch Angriffe abhanden, kann dadurch immenser Schaden entstehen.

Um dieses Risiko zu vermeiden, kann das sogenannte PAUTH-Verfahren (»Pseudonyme Authentifizierung«[3], vgl. Abbildung 1) eingesetzt werden. Dieses Verfahren verwendet eine Kombination aus Kryptographie (sog. »Oblivious Transfer«-Protokoll), Token-Management und Funktionstrennung, um zwei Ziele zu erreichen: Zwei Dienste können im Zusammenspiel eindeutige Pseudonyme zu realen, authentifizierten Nutzern erzeugen, sie sind aber anschließend dennoch nicht in der Lage, die reale Identität zu einem Pseudonym aufzudecken – auch dann nicht, wenn sie sich absprechen oder Daten durch Angriffe oder Leaks abhandenkommen. Allein die Nutzer können die Verbindung bei Bedarf wiederherstellen. Anwendungsfälle für dieses Protokoll werden in ↗**Kapitel 5** diskutiert.

3 vgl. Patent EP 2438707 B1 sowie ↗**https://www.idmt.fraunhofer.de/de/institute/projects-products/privacy-enhancing-technologies.html**.

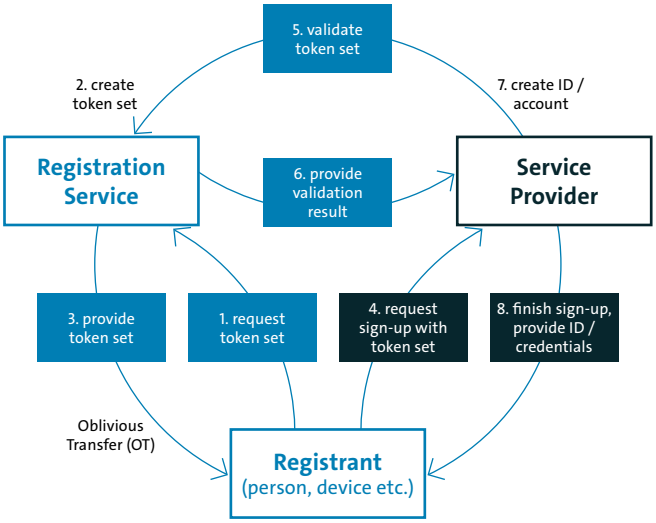

Abbildung 1: Funktionsweise des PAUTH-Verfahrens

2.4 Anonymisierung von Texten

Bei strukturierten Daten kann jedes Attribut nur Werte aus einer stark begrenzten Wertemenge annehmen. Durch dieses Wissen über die möglichen Werte ist eine systematische Anonymisierung möglich.

Häufig sind Daten jedoch semi-strukturiert, d. h. einzelne Attribute eines Datensatzes können beliebig langen natürlichsprachlichen Freitext enthalten. Ebenso liegen Daten häufig in Form von Textdokumenten vor. Hier können Methoden zur Anonymisierung von strukturierten Daten nicht unmittelbar angewendet werden, zumindest nicht, ohne einen hohen Informationsverlust. Dennoch kann man auch sicherstellen, dass natürlichsprachlicher Freitext anonym ist.

Hierbei unterscheiden wir primär drei Möglichkeiten:

Im Voraus Sicherstellen, dass Freitexte keine identifizierenden Begriffe enthalten: Dies ist durch technisch-organisatorische Maßnahmen möglich, z. B. einen eindeutigen Hinweis an dateneingebende Personen.

Nachträgliches Maskieren von identifizierenden Merkmalen: Manuell oder durch Analyseverfahren, die sogenannte Entitäten erkennen, können entsprechende Merkmale extrahiert und entfernt bzw. durch Platzhalter ersetzt werden. Bei kleineren Anonymisierungstätigkeiten auf allgemeinem Text kann sich hier beispielsweise auch ein guter PDF-Editor mit Schwärzungsfunktion als praktisch erweisen.

Neben inhaltlichen Aussagen zu Personen gibt es auch Merkmale, die den Autor eines Textes identifizieren. Daher kann es neben der Anonymisierung des Textinhalts auch von Interesse sein, den Urheber eines Textes anonym zu halten, etwa wenn dieser ein Whistleblower ist, d. h. ein Hinweisgeber auf Missstände. Ein erster Schritt ist es, offensichtliche Autorenangaben, insbesondere in der Titelei von Dokumenten, in der Grußformel von (analogen und elektronischen) Briefen und in diversen Metadaten, zu entfernen.

Ein Autor kann jedoch auch ohne explizite Benennung anhand seines Schreibstils identifiziert werden, sobald geeignete Referenztexte vorliegen. Entsprechende Analysemethoden zur Bestimmung bzw. Überprüfung von Autorschaften gibt es in der Linguistik schon sehr lange. Die automatisierte Autorschaftsanalyse ist hingegen ein recht junges Feld – Pionierarbeiten liegen etwa zwanzig Jahre zurück und erst im vergangenen Jahrzehnt wurde die Forschung hierzu intensiviert. Gute Systeme zur Autorschaftserkennung erreichen bei der Beurteilung, ob ein unbekannter Autor mit einem Referenzautor übereinstimmt, zum Teil eine Genauigkeit von über 80 Prozent.

Spiegelbildlich zu Methoden zur Identifikation von Autoren werden Methoden zur Verschleierung der Autorschaft erforscht.[4] Solche Methoden arbeiten mit Ersetzungen bestimmter Wörter, Paraphrasierung, Umsortierung von Satzteilen oder auch mit Hin- und Rückübersetzungen. Nach heutigem Stand kann die Autorschaft jedoch nicht zuverlässig automatisiert verschleiert werden, wenn der Text lesbar und inhaltlich äquivalent bleiben soll. Derzeit können gute Methoden zur Verschleierung die Chance zur Aufdeckung der Autorschaft in etwa halbieren. Fortschritte bei der Robustheit von Autorschaftserkennungsverfahren können in Zukunft die Verschleierung von Autorschaft allerdings weiter erschweren und den Nutzen aktueller Verschleierungsverfahren weiter schmälern.

Strukturierung mittels Natural Language Processing: Durch Methoden des Natural Language Processing können Freitextdaten strukturiert werden; anschließend können auf diesen strukturierten Daten herkömmliche Methoden zur Anonymisierung von strukturierten Daten angewendet werden. In **↗Kapitel 8** wird ein konkretes Beispiel für die Anonymisierung durch Strukturierung mittels Natural Language Processing am Beispiel medizinischer Freitextdaten gegeben.

Auch bei der Anonymisierung von Freitextdaten hängt es von der Art der zu anonymisierenden Daten, dem geplanten Verwendungszweck der Daten sowie den technischen und organisatorischen Rahmenbedingungen der Datennutzung ab, welches Verfahren anwendbar ist.

4 Martin Potthast, Felix Schremmer, Matthias Hagen, Benno Stein: Overview of the Author Obfuscation Task at PAN 2018:A New Approach to Measuring Safety. In: Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings volume 2125, Sun SITE Central Europe, September 2018, **↗http://ceur-ws.org/Vol-2125/invited_paper_16.pdf**.

2.5 Anonymisierung von Multimedia Daten

Anonymisierung meint im Kontext von Medieninhalten normalerweise das Verbergen von Personenbezügen in Bildern, Videos und Audiodaten, die von Menschen direkt wahrgenommen werden können. Dazu gehören offensichtliche Merkmale wie z. B. Gesichter, Sprachcharakteristika oder Sprachinhalte und textuelle Informationen wie Autokennzeichen oder Namensschilder. Auch weniger offensichtliche Merkmale wie z. B. Körperproportionen oder Gangart, die in der Regel nur ein geübter oder geschulter Beobachter erkennen kann, sind in diesem Zusammenhang zu nennen. Viele der genannten Merkmale werden auch von biometrischen Verfahren genutzt, um Personen zu identifizieren. Im weiteren Sinne personenbezogen können darüber hinaus auch Merkmale sein, die durch Algorithmen und forensische Analysen ermittelt werden können, wie z. B. charakteristische Rauschspuren von Kameras oder eindeutige Aufnahmeprofile von Mikrofonen, mit denen auf Gerätetypen oder Geräte, und damit (indirekt) auf deren Besitzer bzw. Benutzer geschlossen werden kann. Anonymität von Mediendaten ist dementsprechend dann gegeben, wenn aus den Daten weder durch einen Beobachter noch durch biometrische oder andere technische Verfahren ein Personenbezug hergestellt werden kann.

Erreicht wird dies in der Praxis für Bild- und Videomaterial z. B. durch eine Vergröberung (z. B. ein starkes Verpixeln oder Blurring in der Gesichtsregion) oder Substitution (z. B. ein schwarzer Balken über dem Gesicht, siehe Abbildung 2. Für Audio-/Sprachmaterial wiederum bieten sich z. B. Entfernen bzw. Filtern von Sprache, Verfremden der Stimme, Voice Conversion (das »Aufprägen« einer anderen Stimme unter Beibehaltung des Inhalts) oder Sprachsynthese an.

Abbildung 2: Beispielhafte Anonymisierung eines Gesichts (Original links) mittels Verpixelung (mitte) und mittels eines schwarzen Balkens (rechts). Beide Anonymisierungen sind in dieser konkreten Umsetzung ziemlich schwach, da die gezeigte Person noch recht gut zu erkennen ist. (Bildquelle: Pixabay)

In der Praxis wurden und werden derartige Anonymisierungsmethoden allerdings oft nicht mit hinreichender Stärke eingesetzt. Maßstab für eine ausreichende Anonymisierung darf nicht die Frage sein, ob ein durchschnittlicher, unbedarfter Beobachter eine Person erkennen kann. Vielmehr muss man berücksichtigen, dass Personen mit entsprechendem professionellen Hintergrund oder enge Vertraute der betroffenen Person diese überdurchschnittlich gut wiedererkennen können und dass biometrische Algorithmen eine »übermenschliche« Erkennungsfähigkeit haben können. Außerdem sollten, sofern möglich, auch absehbare technische Fortschritte einkalkuliert werden, wie sie für den Bereich des maschinellen Lernens zu erwarten sind (vgl.

Abschnitt 2.6), um entsprechende Sicherheitspuffer bei der Stärke der Anonymisierung einzuplanen.

Ein weiteres Risiko kann auch entstehen, wenn die Originalmedien an anderer Stelle im Internet veröffentlicht werden. In diesem Fall kann etwa über eine inverse Bildersuche ein anonymisiertes Bild mit dem Originalmedium verknüpft und auf diese Weise deanonymisiert werden. Beispielsweise lässt sich das mittlere Bild aus Abbildung 2 über die inverse Bildersuche problemlos finden; die Treffer verweisen dabei auf das Original. Daher sollten Mediendaten, die nach den oben beschriebenen Kriterien anonymisiert sind, tatsächlich als personenbezogene Daten betrachtet werden, wenn bekannt ist, dass diese Medien in nicht-anonymisierter Form im Internet auffindbar sind. Entsprechende Vorsicht ist bei der Weitergabe oder Veröffentlichung der »anonymen« Daten geboten.

2.6 Privatsphärenschutz durch On-Prem-Analyse und Dezentralisierung

Neben Anonymisierung und Pseudonymisierung können auch On-Prem-Analyse und Dezentralisierung zum Schutz der Privatsphäre beitragen, die Datenanalysen unterstützen, ohne eine zentrale Sammlung von personenbezogenen Daten zu erfordern. Diese werden im Folgenden beschrieben.

Beim Einsatz von Software für die Erhebung und Analyse von Daten stehen sich zwei grundsätzlich unterschiedliche Nutzungs- und Lizenzmodelle gegenüber: »On-Premises« (kurz »On-Prem«), d. h. »vor Ort«, und Cloud Computing. Bei »On-Prem« werden Software und Speicherkapazitäten lokal bereitgestellt und gewartet; bei Cloud Computing bzw. »Software as a Service« (SaaS) werden Software, Rechenleistung und Speicherplatz als Dienstleistung von einem externen Anbieter bezogen, bei dem auch die Verantwortung bzgl. Wartung und Betrieb liegt. Es hängt von den konkreten Anforderungen im Einzelfall ab, welches der Modelle das bessere Kosten-Nutzen-Verhältnis bietet, man kann aber pauschal sagen: SaaS bietet vor allem große Vorteile bzgl. Skalierbarkeit (bei Bedarf können Lizenzen bzw. Ressourcen schnell gebucht und oder gekündigt werden), während On-Prem-Lösungen mehr Dezentralisierung, Kontrolle und Eigenverantwortung mit sich bringen. In puncto Datenschutz (und auch bzgl. des Schutzes geschäftsrelevanter Daten) bietet das On-Prem-Modell den Vorteil, dass kritische Daten vor Ort verbleiben und unabhängig von Drittanbietern verarbeitet werden können: Adäquate Sicherheitsmaßnahmen vorausgesetzt, kann ein Zugriff Dritter auf die Daten völlig vermieden werden.

Allerdings kann es aus verschiedenen Gründen schwierig oder gar unmöglich sein, Berechnungen und Datenanalysen lokal auszuführen. Dies ist zum Beispiel der Fall, wenn die lokal vorhandenen Ressourcen bzgl. Hardware, Fachkräften oder Softwarelizenzen nicht ausreichen bzw. die Kosten für zusätzliche Ressourcen zu hoch sind. Ebenso kann es sein, dass Menge und Variabilität der Datenbestände einzelner Akteure nicht ausreichend sind, sodass Daten aus verschiedenen Quellen zusammengeführt werden müssen. Für solche Fälle gibt es Verfahren für sogenannte

homomorphe Verschlüsselung und *sichere Mehrparteienberechnung*, die es erlauben, verteilte Datenbestände gemeinsam zu analysieren und für das Trainieren von KI-Modellen zu verwenden, ohne diese einer zentralen Instanz zugänglich zu machen. Auf diese Weise können die Vorteile von zentralen SaaS-Angeboten mit einem hohen Maß an Dezentralisierung und Kontrolle über die Daten verbunden werden.

Homomorphe Verschlüsselung erlaubt Rechenoperationen auf verschlüsselten Daten. So lassen sich Szenarien realisieren, bei denen Datenbereitsteller schützenswerte Daten – zum Beispiel personenbezogene Daten – zunächst verschlüsseln und dann dem Datenverarbeiter zur Verfügung stellen. Dieser führt anschließend Rechenoperationen auf den verschlüsselten Daten aus. Die Ergebnisse der Berechnungen können aber nur von autorisierten Teilnehmern, z. B. von den Datenbereitstellern, wieder entschlüsselt und verwendet werden. Falls die Analyse so gestaltet war, dass die Ergebnisse nicht mehr personenbezogen sind, können die entschlüsselten Ergebnisse anschließend auch weiterverteilt oder veröffentlicht werden.

Sogenannte vollhomomorphe Verschlüsselung (Fully Homomorphic Encryption) erlaubt prinzipiell beliebige Berechnungen, verursacht aber selbst bei den effizientesten Umsetzungen einen deutlich erhöhten Rechenaufwand, der einem praktischen Einsatz in den meisten Fällen entgegensteht. Einen viel geringeren Mehraufwand bei den Berechnungen verursachen dagegen Verfahren für partiell homomorphe Verschlüsselung (Partially Homomorphic Encryption), die nur bestimmte Rechenoperationen wie z. B. Addition unterstützen, sowie Verfahren, die hinsichtlich der Anzahl der Rechenschritte begrenzt sind (Somewhat Homomorphic Encryption). Derlei spezialisierte Verfahren können für die spezifischen Anforderungen eines Anwendungsfalls ausgewählt und kombiniert werden und eignen sich oft besser für einen Praxiseinsatz, z. B. beim maschinellen Lernen.

Verfahren zur *sicheren Mehrparteienberechnung* (Secure Multi-Party Computation) bzw. sicheren Funktionsauswertung (Secure Function Evaluation) wiederum kommen ganz ohne zentrale Instanz bei der Berechnung aus. Stattdessen tauschen alle beteiligten Parteien verschlüsselte Daten miteinander aus und führen Teilschritte der Berechnung durch, sodass sie am Ende gemeinsam zu dem Ergebnis der Berechnung über alle Eingangsdaten kommen. Für viele Aufgabenstellungen gibt es spezialisierte Protokolle, es gibt aber auch generische Protokolle. Viele Protokolle verwenden »durcheinandergewürfelte Schaltkreise« (Garbled Circuits) oder eine leichtgewichtige Form homomorpher Verschlüsselung. Sichere Mehrparteienberechnung ist in vielen Szenarien praktisch einsetzbar, z. B. bei Wahlen, aber auch beim maschinellen Lernen.

Für den Bereich des maschinellen Lernens wurden in den letzten Jahren Verfahren zum kollaborativen bzw. föderierten Lernen entwickelt. Die Grundidee dahinter ist, gemeinsam KI-Modelle mit den Daten verschiedener Bereitsteller in einem dezentralen Ansatz so zu trainieren, dass Vertraulichkeit und Privatsphäre der Daten geschützt bleiben. Alle Beteiligten führen dazu lokale Trainingsschritte durch und geben lokale Modellinformationen schrittweise an eine zentrale Stelle weiter, die so das Gesamtmodell aktualisiert und dieses wieder zur Verfügung stellt. Erste Ansätze zum föderierten Lernen beruhten auf der Voraussetzung, dass die Eingangsdaten im lokal trainierten Modell bereits hinreichend aggregiert und verschleiert sind. Allerdings wurde

der Informationsgehalt von trainierten Modellen dabei tendenziell unterschätzt (vgl. Abschnitt 2.7). Neuere Ansätze für föderiertes Lernen erhöhen den Schutz der Eingangsdaten mittels Differential Privacy, homomorpher Verschlüsselung, sicherer Mehrparteienberechnung oder einer Kombination derselben. In ↗Kapitel 6 wird föderiertes Lernen ausführlich behandelt und eine konkrete Umsetzung vorgestellt.

2.7 Privatsphärenrisiken beim maschinellen Lernen und Schutzmaßnahmen

Beim maschinellen Lernen kann man zwei Arten von Risiken für die Privatsphäre unterscheiden. Zum einen kann maschinelles Lernen zur Identifikation von Personen genutzt werden und zum anderen kann ein ML-System selbst hinsichtlich der Anonymität der darin enthaltenen Daten untersucht werden. Beide Aspekte werden nachfolgend erläutert und mit Beispielen versehen.

In Bezug auf den ersten Aspekt, also der Identifikation von Personen durch ML-Systeme, haben Fortschritte bei den ML-Verfahren in Verbindung mit Big-Data-Technologien die Grenzen der praktischen Anonymität verschoben. Die technischen Fortschritte ermöglichten in den vergangenen Jahren die Erschließung von immer mehr Datenquellen, die zuvor aufgrund ihrer Art und ihres Umfangs nicht mit akzeptablem Aufwand automatisierten Analysen unterzogen werden konnten. Daher können heute Personen auch in einer Flut an unstrukturierten Daten identifiziert werden. Beispielsweise wird die Gesichtserkennung mittlerweile von mehreren Staaten und Dienstleistern in einem Maßstab praktiziert, der etwa im Jahr 2015 nur für wenige Pioniere ein vorstellbares Ziel war.

Ebenso ermöglicht maschinelles Lernen das Deanonymisieren von Daten, die zuvor vermeintlicherweise als anonym galten. Beispielsweise können verpixelte oder weichgezeichnete Gesichter oder Kfz-Kennzeichen mit Hilfe neuronaler Netze besser erkannt, rekonstruiert oder entziffert werden als dies für einen menschlichen Betrachter möglich ist[5], vgl. Abschnitt 2.5. Ebenso können ML-Systeme den Verfasser eines Textes anhand sprachlicher Merkmale, welche die Diktion eines Textes und damit Aspekte des persönlichen Stils eines Autors erfassen, recht zuverlässig identifizieren, vgl. Abschnitt 2.4.

Maschinelles Lernen sollte in Anbetracht solcher Möglichkeiten zur Deanonymisierung nicht als das ursächliche Problem betrachtet werden, sondern als Werkzeug, um Risiken bei der Anonymisierung zu erkennen. Dies muss mit einer gründlicheren Anonymisierung der Daten gelöst werden. Maßgeblich ist hier die Frage, was prinzipiell an personenbezogener Information in den Daten nach der Anonymisierung verbleibt. Es reicht nicht als Bewertungskriterium zu fordern, dass ein menschlicher Betrachter Personen nicht mehr erkennen kann. Dahingegen ist bei der Identifikation von Personen in nicht-anonymisierten Massendaten mittels maschinellem Lernen

5 Richard McPherson, Reza Shokri, Vitaly Shmatikov: Defeating Image Obfuscation with Deep Learning. Computing Research Repository (CoRR), Article ID arXiv:1609.00408v2 [cs.CR], arXiv, September 2016.

jedoch die Nutzung des maschinellen Lernens selbst sowie der Zugriff auf die jeweiligen Datenquellen im konkreten Anwendungsfall rechtlich, politisch und ethisch zu bewerten.

Das zweite Risiko, das in diesem Abschnitt beleuchtet werden soll, ist die Anonymität von ML-Systemen selbst bzw. genauer gesagt die Anonymität von den darin gespeicherten Daten. Lange wurde in der Praxis die Ansicht vertreten, dass durch das Training eines neuronalen Netzes die Daten in einem ML-System so sehr abstrahiert und aggregiert werden, dass parallel eine vollständige Anonymisierung des zugrundeliegenden Datenmaterials geschieht. Aktuelle Forschungen zeigen jedoch, dass hier das Risiko besteht, dass eine unerwartet klare Erinnerung an die Trainingsdaten im neuronalen Netz verbleibt. Diese kann von Angreifern genutzt werden, um Rückschlüsse über die Trainingsdaten zu ziehen oder gar die ursprünglichen Trainingsdaten annähernd zu rekonstruieren und somit die Privatheit der Datensubjekte zu gefährden.

Dabei wurde zunächst aufgedeckt, dass Systeme, die zum Generieren synthetischer Daten nach dem Vorbild realer Daten genutzt werden, durchaus Stücke von Trainingsdaten mit einer privatsphärerelevanten Größe und Häufigkeit wiedergeben. Das heißt, dass solche Artefakte hinreichend groß sind, dass sie individuelle Merkmale oder Merkmalskombinationen von Personen aus den Eingangsdaten wiedergeben, und dass solche Artefakte weit häufiger auftreten, als durch eine zufällige Generierung aus einer adäquaten Wahrscheinlichkeitsverteilung zu erwarten wäre. Beispielsweise können neuronale Netze Kreditkartennummern aus den Trainingsdaten preisgegeben.[6]

Andere Forschungsansätze zielen darauf ab, auch bei solchen ML-Systemen Rückschlüsse auf die verwendeten Trainingsdaten zu ziehen, bei denen das Modell nicht dazu genutzt werden kann, Ausgabedaten nach dem Vorbild der Trainingsdaten zu generieren. So ist es bei ML-Systemen teilweise möglich, Rückschlüsse zu ziehen, ob ein konkretes Testdatum in den Trainingsdaten enthalten war (Membership Inference).[7] Wann immer ein solches System auf Daten zu Einzelpersonen trainiert wurde, ist somit die Privatsphäre dieser Personen gefährdet. Wenn ein System beispielsweise darauf trainiert wurde, für Menschen mit einer bestimmten Erkrankung Empfehlungen bzgl. der Wahl der Behandlung zu geben, dann offenbart die Zuordnung einer bestimmten Person zur Trainingsmenge, dass die überprüfte Person die Erkrankung hat.

Ein ähnliches Angriffsziel ist es, Trainingsdaten eines ML-Systems zu rekonstruieren (Model Inversion).[8] Auch hier sind die Personen, die Trainingsdaten gestellt haben, dem Risiko ausgesetzt, von Angreifern bestimmten, evtl. stigmatisierenden, Merkmalen zugeordnet zu werden.

6 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song: The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. Computing Research Repository (CoRR), Article ID arXiv:1802.08232v3 [cs.LG], arXiv, Juli 2019.

7 Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov: Membership Inference Attacks Against Machine Learning Models. In: IEEE Symposium on Security and Privacy 2017. Seiten 3–18, 2017.

8 Matt Fredrikson, Somesh Jha, Thomas Ristenpart: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: ACM Conference on Computer and Communications Security 2015, Seiten 1322–1333, 2015.

Eine ausführliche Darstellung verschiedener Angriffsmöglichkeiten auf neuronale Netze wird in ↗**Kapitel 7** mit einem Schwerpunkt auf Bilddaten gegeben.

Um Privatsphärerisiken von ML-Modellen abzuwenden, kann man verschiedene Schutzstrategien in den unterschiedlichen Phasen des maschinellen Lernens einsetzen. Zunächst können die Trainingsdaten selbst anonymisiert werden. Hierbei ist das Hinzufügen von Rauschen mittels Mechanismen für Differential Privacy die Strategie der Wahl, da Anonymisierungen mittels Generalisierungs- und Aggregationsstrategien Verteilungsartefakte erzeugen, die zu ungeeigneten ML-Modellen führen können. In der Phase des Trainingsprozesses stehen die Strategien des föderierten Lernens zur Verfügung, welche sich ebenfalls Differential Privacy oder auch homomorphe Verschlüsselung und Sichere Mehrparteienberechnung zunutze machen können (vgl. Abschnitt 2.6). Schließlich können auch in der Nutzungsphase des Modells die Ausgaben gegen Privatsphärerisiken geschützt werden, was das primäre Ziel ist, sofern das Modell selbst nicht weitergegeben, sondern in einer sicheren Umgebung betrieben wird. Falls in den vorhergehenden Phasen geeignete Schutzmaßnahmen ergriffen wurden, sind die Ausgaben bereits implizit geschützt. Andernfalls können die Ausgabedaten durch Generalisierung oder auch durch Differential Privacy geschützt werden. Alle Schutzmaßnahmen zielen letztlich darauf ab, durch eine Ungenauigkeit oder einen Fehler vorgegebener Stärke die Rekonstruktion der Trainingsdaten in einer die Privatheit gefährdenden Qualität zu verhindern. ↗**Kapitel 7** erläutert verschiedene Schutzmaßnahmen in größerer Tiefe.

3 Speicherung von Geo-Bewegungsprofilen

3 Speicherung von Geo-Bewegungsprofilen

Michael Mundt

Aufenthaltsort natürlicher Personen, Bewegung von Personen über einen substantiellen Zeitraum, Erstellung von Profilen über die Bewegung

In der klassischen Betrachtungsweise von Geodaten liegt kein Bezug zu einer natürlichen Person vor. Straßen, Oberflächenmodelle, Gebäude, Flüsse und Nutzungsflächen beispielsweise weisen keinen Bezug zu einer natürlichen Person auf. Es liegt keine im Datenschutz begründete Notwendigkeit vor, diese Daten zu pseudonymisieren oder zu anonymisieren. Jedoch ist ein Trend zu verzeichnen. Im Zuge der Digitalisierung werden zunehmend Sensoren eingesetzt, die im höheren Detail auch den Standort und die Zeit des Datums aufzeichnen. Dies geschieht heute in hoher Genauigkeit. Der Standort wird präzise in der Lage aufgenommen. Ein prägnantes Beispiel dafür sind heutige Mobiltelefone deren Position z. B. über Verschneidungen der IP-Adressen in Kombination mit der Laufzeit des Versands bestimmt werden kann. Zumeist sind Mobiltelefone heute auch mit eigenen Positionssensoren wie GPS ausgestattet und können selbst den aktuellen Aufenthaltsort abrufen. Es ist ferner davon auszugehen, dass z. B. mobile Endgeräte über gerätespezifische Kennzeichnungen der nutzenden Person oder zumindest dem Besitzer zugeordnet werden können. Folgerichtig ist in der Datenschutzgrundverordnung (DSGVO) dieser Aspekt berücksichtigt worden:

DSGVO, Art. 4 Begriffsbestimmung Abs. 1:

»›personenbezogene Daten‹ bezeichnen alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder [...], identifiziert werden kann.«

DSGVO, EG (26):

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

Die Kennungen Standortdaten und Kennnummer sind benannt. Standortdaten bezeichnen den tatsächlichen Aufenthaltsort einer Person, bzw. vergangene oder prädizierte Aufenthaltsorte.

Unter Kennnummer ist z. B. ein gerätespezifischer Bezeichner zu verstehen. Es kann statuiert werden, dass Geodaten mit hoher räumlicher Genauigkeit im Sinne der DSGVO daraufhin zu überprüfen sind, ob über den Standort und die Kennung des digitalen Aufzeichnungsgerätes eine natürliche Person identifiziert werden kann. Dabei können Mittel allgemeinen Ermessens unterstützend verwendet werden. In diesem Falle handelt es sich um personenbezogene Daten.

Dies trifft gewiss für Bewegungsprofile zu. Die folgende Grafik zeigt exemplarisch die Aufzeichnung eines Bewegungsprofiles. Zur Erfassung der Daten wurde eine Anwendung (»App«) auf einem Smartphone verwendet.

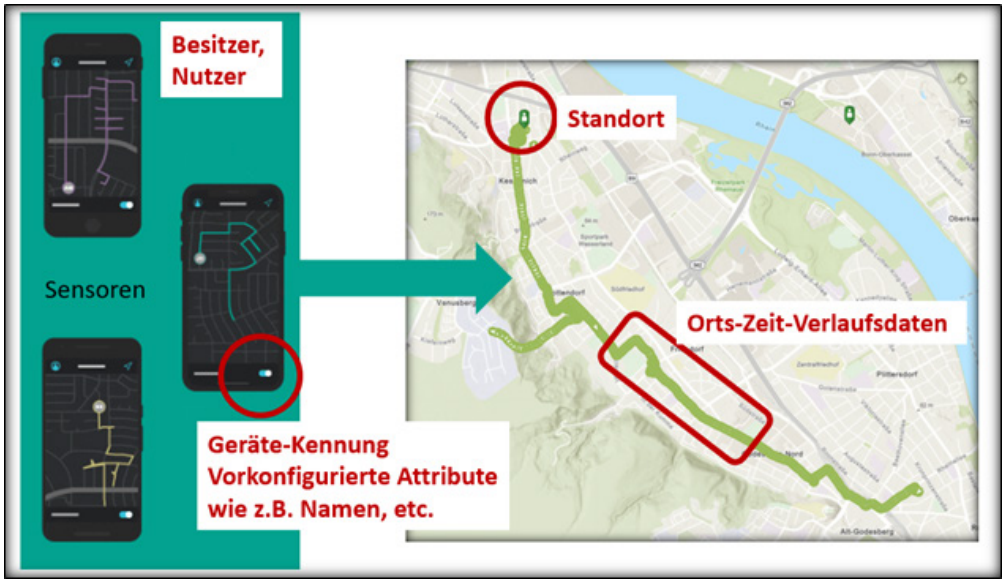

Abbildung 3: Bewegungsprofil, aufgezeichnet mit einem Smartphone

Es ist leicht vorstellbar, wie mit geringem Aufwand an Zeit und Kosten und nach aktuellem Stand der Technik ein solches Bewegungsprofil z. B. mit Adressen auf der Karte verschnitten werden kann. Wir zusätzlich noch der zeitliche Verlauf berücksichtigt, so ist die natürliche Person identifizierbar. Wenn jeden Abend an Wochentagen nach ca. 17 Uhr der Standort das Geo-Bewegungsprofil an einer bestimmten Adresse angezeigt wird, dann handelt es sich mit hoher Wahrscheinlichkeit um den Wohnort der Person. Ein Blick in das Adressregister identifiziert dann die Person. Wenn es möglich ist, z. B. auf Basis einer Gerätekennung des aufnehmenden Gerätes Muster in den Geo-Bewegungsprofilen zu erkennen, dann können Rückschlüsse auf den Kontext der Person gezogen werden. Beispielsweise die Reisetätigkeit zu Arbeitszeiten, Freizeitverhalten oder auch Arztbesuche laufen Gefahr, offengelegt zu werden. Mitunter senden Anwendungen mobiler Endgeräte wie Smartphones vorkonfigurierte Profile mit den Daten. Hierbei kann es sich um Angaben zur Person handeln. In diesem Falle ist es direkt möglich, das

Geo-Bewegungsprofil einer natürlichen Person oder zumindest dem derzeitigen Nutzer des Gerätes zuzuordnen.

Im Falle des Geo-Bewegungsprofiles sind nun die Maßnahmen des Datenschutzes umzusetzen. Folgende Maßnahmen bieten sich an:

- Im Sinne datenschutzfreundlicher Standardeinstellungen sind die Anwendungen (»Apps«) a priori so zu konfigurieren, dass keine Attributinformationen zum Besitzer/ Nutzer versendet werden (Bezug: DSGVO Art. 25 Datenschutz durch Technikgestaltung und datenschutzfreundliche Voreinstellungen)

- Der Zweck der Erfassung des Geo-Bewegungsprofils ist festzuhalten, die Daten sind geeignet gegen Offenlegung zu schützen sowie der Zugriff auf die Daten einzuschränken (Bezug: DSGVO Art. 32 Sicherheit der Verarbeitung)

- Klare, organisatorische Anweisung zu den Aufzeichnungszeiten der Geo-Bewegungsprofile; strikte Trennung zwischen dienstlichen und privaten Zeiträumen

- Frühestmögliche Pseudonymisierung der Gerätekennungen oder sogar – sofern es der Verarbeitungszweck erlaubt – die Anonymisierung der Gerätekennungen (Bezug: DSGVO Art. 32 Sicherheit der Verarbeitung)

- Durch Aggregation der Orts- und Zeitinformationen (Bezug: Standort) ist eine Trennung der personenbezogenen Daten zu erwirken, soweit es der Verarbeitungszweck erlaubt

Der Vorgang der Aggregation ist spezifisch für Geo-Bewegungsprofile umzusetzen. Hierzu kommen räumlich-analytische Funktionen zum Einsatz. Beispielsweise kann das Geo-Bewegungsprofil mit dem Straßennetz analytisch verschnitten werden. Hieraus kann dann z. B. die Nutzung der Straßen aufaddiert werden. Mehrere Bewegungsprofile können zu Gruppen zusammengefasst werden. Analytisch können die gefahrenen Kilometer ermittelt und dann z. B. in Form eines räumlichen Schwerpunktes (z. B. Centroid oder unregelmäßige Ausdehnungsfläche) als Datensatz verortet werden. In Abhängigkeit des Verarbeitungszwecks stehen derart vielfältige Aggregations-Möglichkeiten zur Auswahl. Dabei ist sicherzustellen, dass der Bezug zur natürlichen Person zum Beispiel über eine private Adresse nicht mehr rekonstruiert werden kann.

Situationsbedingt kommt dieses Verfahren der Aggregation derzeit zur Lagedarstellung der Ausbreitung des Coronavirus zum Einsatz. Bestätigte Infektionen werden aggregiert und den Landkreisen zugeordnet. Auf diese Weise wird der Bezug zur Person separiert und eine anonymisierte Information bereitgestellt. Im Sinne des Datenschutzes wird auf diese Weise eine Information zum Zwecke eines Lagebildes verarbeitet und dabei werden unbedingt die Persönlichkeitsrechte der betroffenen Personen geschützt. Auf Basis der Anzahl der Infektionen auf Landkreisebene ist es nicht mehr möglich, auf die einzelne Person und deren Standort Rückschlüsse zu ziehen.

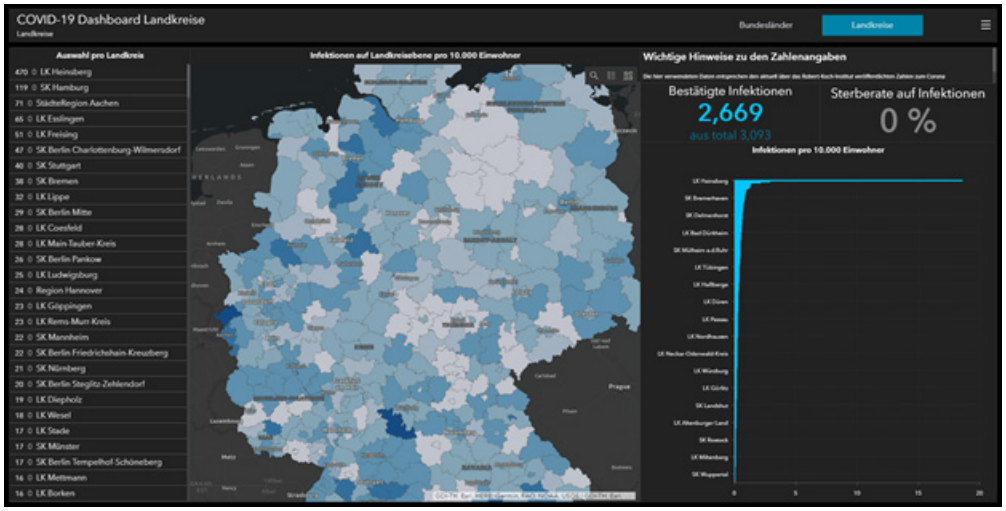

Abbildung 4: COVID-19 Dashboard des Robert Koch Institutes

4 Use Case: Google's COVID-19 Community Mobility Reports

4 Use Case: Google's COVID-19 Community Mobility Reports

Christoph Dibak, Vadym Doroshenko, Yurii Sushko

Differential Privacy, Google Safety Engineering Center, Open Source

4.1 Introduction

Leveraging user data allows Google to build great services and provide valuable insights to the community. When doing so, it is crucial to ensure the privacy of each individual. How to quantify privacy? This question is not only philosophical but also technical. One approach for defining privacy in a technical and mathematically measurable way is so-called »differential privacy«. Simply explained, differential privacy allows limiting the effect of an individual's contributions on the final output.[9]

Google uses differential privacy to enable services that rely on user's data in a privacy-preserving manner. One very recent and highly visible use case of differential privacy is Google's COVID-19 mobility reports, which provide insights into the changes in people's mobility patterns during the epidemics[10]. For the launch of mobility reports, several Google anonymization and privacy experts from the Google Safety Engineering Center in Munich worked together to ensure the users' privacy is respected. One very recent and highly visible use case of differential privacy is Google's COVID-19 mobility reports (see Abb. 5)

Abbildung 5: Screenshot of the COVID-19 mobility reports taken on April 17 2020 shows changes in visits to different classes of places in Germany during the pandemic based on anonymized data.

9 Cynthia Dwork and Aaron Roth (2014), »The Algorithmic Foundations of Differential Privacy«, Foundations and Trends® in Theoretical Computer Science: Vol. 9: No. 3–4, pp 211-407. ↗http://dx.doi.org/10.1561/0400000042.

10 The mobility reports are available online ↗https://www.google.com/covid19/mobility/.

Have the visits to public places (groceries, parks, recreational places) become less frequent after the lockdown? If so, to what extent? Publication of mobility reports allows us to answer those questions and provides a helpful resource for researchers and decision makers.

As COVID-19 mobility reports rely on highly sensitive location data, it is of crucial importance to ensure the privacy of all individuals. In the following section, we illustrate how differential privacy is applied to achieve that goal.

4.2 Data Anonymization Strategy

This publication only provides a high-level overview of the anonymization strategy for the COVID-19 mobility reports. In particular, we will focus on the aggregation and anonymization of the grocery store visit statistics. Full details about the anonymization strategy for the mobility reports has been published in a separate publication.[11]

Mobility reports use aggregated and anonymized data from the Google Location History. Location History is off by default and users have full control over their data by using the Activity Controls[12] of the Google Account to enable or disable all location history, by using the Google Maps Timeline[13] to delete individual visits, or by temporarily activating incognito mode in Google Maps. Users can choose to enable this feature to, e.g., see recent visits in Google Maps.

The computation of time spent by users in grocery stores is based on an anonymized aggregation that is performed per day and geographic regions, e.g., on a country level and on a state level in Germany. In particular, we are using a differential private mean mechanism. This mechanism adds random Laplace noise[14] to the true aggregate of grocery store visits depending on the location of grocery stores available in Google Maps. Adding Laplace noise ensures the privacy of each individual. Additionally, each user is counted in no more than four pairs of <category, location> pairs per day, e.g., if a user went to grocery stores, parks, and retail places in Berlin and Potsdam, which are nearby cities in different states in Germany, on a single day, only four of those pairs are selected randomly and utilized for the aggregation. This allows limiting the contribution of each user which therefore requires less noise and consequently having more utility while still providing the same privacy guarantees.

Once the data is anonymized, we can perform any post-aggregation steps and combine it with other anonymized data. In particular, we are reporting changes over a baseline taken from a time period before the pandemic started in most parts of the world. The baseline is aggregated and anonymized using the same steps as above. Additionally, we use an anonymized count to

11 Aktay et al. Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.0), April 2020, available online ↗**https://arxiv.org/abs/2004.04145**.

12 ↗**https://myaccount.google.com/activitycontrols**

13 ↗**https://maps.google.com/timeline**

14 ↗**https://en.wikipedia.org/wiki/Laplace_distribution**

replace any regions where we do not have enough data and where the data might only have low significance.

4.3 Open Source Library

The data anonymization process for the COVID-19 mobility reports is implemented using Google's differential privacy library. Like for cryptography, there is a lot of room for conceptual and technical pitfalls when implementing anonymization, that take time to detect and fix.[15] Google's engineers have worked on differential privacy since 2014 and have spent a lot of time on implementing, hardening, and testing this library. To help the industry and the scientific community use differential privacy and safely implement anonymization in their projects, Google open-sourced this library and made it publicly available for everyone[16]. Engineers from the Google Safety Engineering Center in Munich and other offices are continuing to work on open source solutions for developers and organizations to use differentially-private data analysis.

4.4 Summary

Google uses differential privacy for releasing the COVID-19 mobility reports as a resource for public health authorities during the coronavirus crisis. We discussed how data is anonymized using the example of how the number of visits to grocery stores has changed. As developing differential privacy is complex and has many technical difficulties to make it right, Google has open-sourced their underlying differential privacy library, ready to be used by other companies and for other projects.

15 Mironov, Ilya. »On significance of the least significant bits for differential privacy.« Proceedings of the 2012 ACM conference on Computer and communications security. 2012.

16 Google's differential privacy library is available at ↗**https://github.com/google/differential-privacy**

5 Anwendungsfälle für »entkoppelte Pseudonyme«

5 Anwendungsfälle für »entkoppelte Pseudonyme«

Steffen Holly, Patrick Aichroth

Anonymisierung von Fahrzeugdaten, Datenaustausch mit entkoppelten Pseudonymen, Authentifizierung ohne Identifizierung

Datenaustausch ist die Basis für Mehrwerte und neue Geschäftsmodelle. Aber das Risiko, kritische Informationen preiszugeben – im schlimmsten Fälle an potenzielle Konkurrenten – oder gegen rechtliche Vorgaben zu verstoßen, verhindert in vielen Fällen einen solchen Austausch. Dahinter steht ein Dilemma zwischen dem Schutz von Datenquellen bzw. Datenschutz einerseits, und dem Bedarf nach einer differenzierten Datenanalyse andererseits: Der Einsatz von Pseudonymisierung erlaubt zwar eine differenzierte Datenanalyse, birgt aber erhebliche Risiken, dass reale Identitäten und Datenquellen im Ernstfall wieder aufgelöst werden können – besonders dann, wenn über eine Authentifizierung gewährleistet sein soll, dass hinter den entsprechenden Pseudonymen reale Nutzer oder andere reale Entitäten stehen. Anonymisierung wiederum kann dieses Problem zwar vermeiden, erlaubt aber keine differenzierten Analysen mehr, und ist daher für solche Anwendungsfälle ungeeignet.

Eine Möglichkeit zur Lösung dieses Dilemmas ist in ↗**Kapitel 2.3** mit der Entkopplung von Pseudonymen beschrieben. Das Protokoll[17] liefert eindeutige IDs bzw. Pseudonyme für die Datenanalyse, eine etwaige Rückführung der IDs auf reale Identitäten durch Dritte wird aber im Gegensatz zu herkömmlichen Pseudonymisierungsverfahren verhindert – das gilt auch dann, wenn sich Dritte absprechen oder Daten unabsichtlich bzw. vorsätzlich abhandenkommen. Auf diese Weise werden schwierige Datenschutz- und Sicherheitsprobleme bereits im Ansatz vermieden, die reale Identität des Datenbereitsteller bleibt geschützt.

So können Anwendungsfälle realisiert werden, bei denen reale Nutzer, Geräte, Dinge oder Organisationen einen vorab definierten Authentifizierungsprozess durchlaufen und anschließend differenzierte Datenanalysen nach IDs möglich sein sollen, gleichzeitig aber die Datenherkunft und damit verbundene sensible Informationen außen vor bleiben sollen. Damit lassen sich Privatsphäre und Datenschutz bei der Speicherung und Verarbeitung von personenbezogenen Daten (in Apps, Anwendungen), datenschutzfreundliche Personalisierungs- und Empfehlungssysteme, Datenaustausch zwischen Partnern und Konkurrenten in Wertschöpfungsketten, kollaborative verkettete Routenplanung, B2B Benchmarking, datenschutzfreundliche Umfragen und andere Fälle realisieren, die normalerweise an dem beschriebenen Widerspruch zwischen funktionalen Anforderungen bzgl. Datenanalyse und Authentifizierung von Akteuren oder Objekten einerseits und den Sicherheits- und Datenschutzanforderungen andererseits scheitern würden.

17 White Paper Pseudonyme Authentifizierung ↗**www.psoido.com**

5.1 Privatsphäre und differenzierte Datenanalysen für Fahrzeugdaten

Um vernetzte Services mit Fahrzeugen zu nutzen, muss der Besitzer des Fahrzeuges sein Einverständnis gegenüber dem Automobilhersteller geben, wobei i.d.R. auch die Nutzung der sonstigen Fahrzeugdaten und Sensoren für die Analyse beim Hersteller erlaubt wird. Verarbeitung und Speicherung dieser Daten erfolgt dann beim Hersteller oder in dessen Auftrag entsprechend der DSGVO. Das gilt auch bei der Verwendung von Pseudonymen für differenzierte Datenanalysen, und bei jeder neuen Verwendungsform ist ein explizites Einverständnis des Nutzers erforderlich. Parallel dazu werden Daten z. B. von Navigationssystemen aggregiert und anonymisiert, um sie für die Verbesserung von Karten und Diensten zu nutzen. Aufgrund der Anonymisierung ist ein Einverständnis des Nutzers hierfür nicht erforderlich, die Daten sind allerdings auch nicht mehr differenzierbar und auch nicht für Personalisierungen verwendbar.

Die Verwendung von »entkoppelten Identitäten« kann in diesem Zusammenhang erhebliche Vorteile mit sich bringen: Nach der Authentifizierung von Nutzern könnte automatisch und quasi im Hintergrund eine neue, nicht mehr auf die reale Identität rückführbare ID ausgegeben werden (vgl. ↗Kapitel 2.3). Diese könnte dann zur Speicherung von Nutzerdaten wie z. B. Beschleunigung, Geschwindigkeit im Verhältnis zur erlaubten Geschwindigkeit (ohne GPS-Koordinaten), Benzinverbrauch, Drehzahlen, Gewicht der Zuladung u.ä. verwendet werden. Dabei muss allerdings darauf geachtet werden, dass diese Daten nicht per se zu einer Re-Identifizierung der betreffenden Fahrer führen können. Dazu können verschiedene Standardverfahren der Anonymisierung (vgl. ↗Kapitel 2) eingesetzt werden, um das Verfahren zur Erzeugung »entkoppelter Identitäten«) zu flankieren.

Der so entstehende Datenpool bestehend aus IDs und zugehörigen Daten enthält dann differenzierbare Profile, die durch niemanden als den Nutzer selbst auf dessen reale Identität zurückgeführt werden können. Im Ergebnis liefert dieser Ansatz eine Lösung, die eine Stärkung des Datenschutzes erstmalig mit den Möglichkeiten differenzierter Analyse bis hin zu personalisierten Diensten verbindet.

5.2 Datenaustausch ohne preisgabe kritischer Informationen

Geradezu prädestiniert ist das Protokoll der pseudonymen Authentifizierung für Bereiche, bei denen der Austausch von Daten zwischen Partnern oder Wettbewerbern nötig ist, um die Effizienz im System zu erhöhen oder um neue Mehrwerte zu erschließen. Denn hier besteht oft die Herausforderung, dass die Teilnehmer nicht zur Bereitstellung der Daten bereit sind, weil sie Sorge haben, auf diese Weise Daten preiszugeben, die Rückschlüsse auf geschäftskritische Abläufe geben könnten. Auf der anderen Seite ist ein Austausch aber oft unbedingte Voraussetzung zur Erschließung von Effizienzpotenzialen – ein Dilemma.

Hierzu ein Beispiel: Ein Maschinenhersteller verkauft seine vernetzten Maschinen an Kunden weltweit und möchte gerne die bei den Kunden erzeugten Maschinendaten analysieren, um Schwierigkeiten und Verbesserungspotenziale zu identifizieren und die Produkte zu verbessern. Obwohl diese Produktverbesserungen auch den Kunden und Nutzern der Maschinen zugute kommen, möchten diese die Maschinendaten aber nicht teilen, weil sie damit Daten über deren Nutzung und z. B. Produktionsauslastung des Unternehmens preisgeben würden (die im schlimmsten Fall sogar in den Händen von Konkurrenten landen könnten).

Durch den Einsatz des beschriebenen Verfahrens zur Entkopplung von Identitäten kann dieses Dilemma aufgelöst werden: Durch die Erzeugung von nicht rückführbaren, aber dennoch eindeutigen (und authentifizierten) Maschinen-IDs können Maschinendaten übertragen werden, ohne dass Rückschlüsse auf deren Herkunft möglich wären. Kritische Daten verbleiben beim Kunden, der Hersteller kann aber dennoch differenzierte Analysen auf Basis eindeutiger Profile durchführen und individuelle Empfehlungen zu Wartung anonym übertragen.

5.3 Mehrwerte von entkoppelten Identitäten durch pseudonyme Authentifizierung

Durch die Entkopplung von Identitäten werden Anbieter in die Lage versetzt, zwei oft widersprüchliche Anforderungen miteinander zu vereinbaren: Die Anforderung nach einem starken Datenschutz, und die Möglichkeit zu einer differenzierten Analyse und möglichst flexiblen Nutzung der Daten, bis hin zu personalisierten Diensten.

Mit potenziell nicht mehr zustimmungspflichtigen Daten haben z. B. Automobilhersteller die Möglichkeit, Daten mit Dritten zu teilen oder Versicherungen zur Verfügung zu stellen. Durch die erhaltene Differenzierung von Identitäten sind bessere Analysen möglich, und dennoch bleiben die Datenschutzanforderungen gewahrt. Der Nutzer kann von maßgeschneiderten Angeboten des Autoherstellers oder dessen Partnern profitieren, und bleibt trotzdem für Partner und Hersteller anonym.

Maschinenhersteller wiederum können über die so erhobenen Daten eine stärkere Kundenbindung durch Produktupdates und verbesserte Angebote, aber auch pay-per-use Geschäftsmodelle auf Basis der anonymen Datenprofile realisieren, ohne dadurch die legitimen Vertraulichkeitsbedürfnisse ihrer Kunden einzuschränken. Die Kunden wiederum profitieren von individuellen, auf tatsächlichen Nutzung basierenden Updates und Produktverbesserungen, und damit von einer zuverlässigen Produktion, geringeren Standzeiten und mehr Effizienz.

6 Föderiertes Lernen: Bringt die Algorithmen zu den Daten statt die Daten zu den Algorithmen

6 Föderiertes Lernen: Bringt die Algorithmen zu den Daten statt die Daten zu den Algorithmen

Michael Huth & Markus Kaulartz

Kollaboratives Maschinelles Lernen, Pseudonymisierte KI Modelle, Anonymisierte KI Modelle, Kryptographische Protokolle zur Wahrung der Privatsphäre, Personenbezug von kollaborativ errechneten KI Modellen

Ähnlich wie in der Politik ist der föderale Ansatz das Gegenstück zum Zentralismus. Beim föderierten Lernen müssen keine Rohdaten zu zentral laufenden Algorithmen bewegt werden, sondern die Algorithmen laufen in der Umgebung der Rohdaten. Dies erleichtert das Einhalten von Datenschutzanforderungen erheblich ohne die Effizienz des maschinellen Lernens zu beeinträchtigen.

Trainieren eines KI-Modells auf der Basis umfangreicher Daten ist ein essentieller Bestandteil des maschinellen Lernens. Dabei gilt: Je mehr Daten verfügbar sind, desto besser bzw. genauer ist das resultierende KI-Modell.

Die notwendigen Daten liegen in der Regel nicht bereits aggregiert an dem für das Training vorgesehenen Ort vor, sondern werden über eine Vielzahl von Quellen – z. B. Datenbanken, Smartphones oder IoT-Devices – gewonnen bzw. gespeichert. Sie sind beispielsweise getrennt nach Kunden in unterschiedlichen Unternehmensdatenbanken abgelegt, auf Endgeräten wie etwa autonomen Autos gespeichert oder werden mithilfe von smarten Voice-Assistenten generiert. In all diesen und vergleichbaren Fällen könnte eine Aggregierung der Daten nicht nur technisch schwierig und kostenintensiv sein, sondern den Anforderungen der DSGVO widersprechen und damit gegen geltendes Recht verstoßen.

Grundsätzlich ist großvolumiges Aggregieren bzw. zentrales Sammeln sensibler oder personenbezogener Daten durch Unternehmen mit enormen Risiken behaftet oder gar widerrechtlich – angefangen vom unbefugten Zugriff auf diese Daten bis zu deren Weiterverarbeitung in rechtlichen Grauzonen. In den letzten Jahren haben sich mehr Kunden für den Datenschutz sensibilisiert, sodass dieser Ansatz negative Auswirkungen auf Umsätze und das jeweilige Branding von Unternehmen haben kann.

Eine technische Lösung für dieses Problem ist das föderierte Lernen. Dieser methodische Ansatz kann Beschränkungen und Risiken von zentralistisch konzipierten Datenauswertungsmethoden in wertschöpfende Möglichkeiten umwandeln. Beim föderierten Lernen werden KI-Modelle trainiert, ohne dass die Daten ihren Ursprungs- bzw. Speicherort verlassen und ohne, dass der Verantwortliche Zugriff auf die rohen Trainingsdaten bekommt. Dieses dezentrale Konzept ermöglicht in einer Vielzahl von Szenarien überhaupt erst gesetzeskonformes maschinelles Lernen.

6.1 Was ist föderiertes Lernen?

Beim föderierten Lernen bleiben die Daten in den Umgebungen, in denen sie entstehen und gespeichert sind. Das Trainieren erfolgt ausschließlich in diesen lokalen Umgebungen und produziert ein lokales KI-Modell mittels seiner lokalen Daten. Diese lokalen KI-Modelle werden an einen Koordinator geschickt, der sie in ein neues globales KI-Modell aggregiert. Das sich daraus ergebende, globale KI-Modell kombiniert das implizite Wissen aller lokalen KI-Modelle und kann so genauer als jedes lokale KI-Modell sein.

Nach der Aggregierung wird dieses neue globale KI-Modell an die lokalen Umgebungen verteilt, um erneut zu lernen (siehe Abbildung 6).

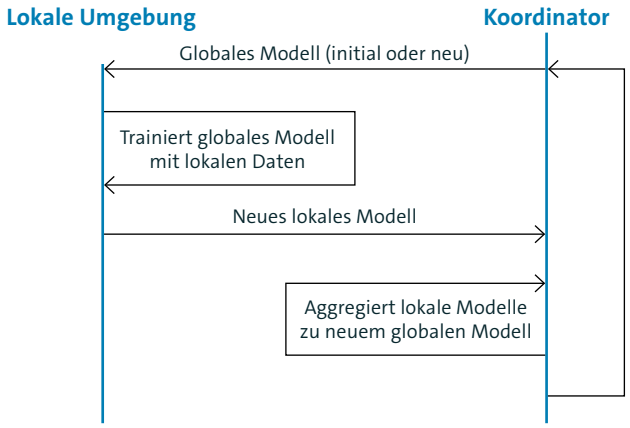

Abbildung 6: Schematische Darstellung einer Runde des föderierten Lernens. Zuerst sendet der Koordinator an lokale Umgebungen ein neues globales KI-Modell. Anschließend trainieren die lokalen Umgebungen mit lokalen Daten dieses erhaltene globale Modell. Die daraus resultierenden neuen lokalen Modelle werden jeweils an den Koordinator geschickt. Der Koordinator aggregiert die erhaltenen lokalen Modelle zu einem neuen globalen Modell und schickt es an lokale Umgebungen, um eine weitere Runden des föderierten Lernens einzuleiten.

Dies beendet eine *Runde* des föderierten Lernens. In der nächsten Runde wird nun das globale Modell durch weiteres Trainieren mit lokalen Daten zu einem neuen lokalen Modell verändert. Diese neuen lokalen Modelle werden wie in der vorherigen Runde an den Koordinator geschickt, welcher sie erneut in ein neues globales Modell aggregiert. In der Praxis können mehrere hundert oder tausend Runden ausgeführt werden bis ein globales Modell mit gewünschter Genauigkeit oder zusätzlichen Eigenschaften berechnet wird.

6.2 Anwendungsbeispiel zum föderierten Lernen

Ein global operierendes Unternehmen mit circa 200.000 Mitarbeitern möchte einen Sprachassistenten entwickeln, der jeder/m Mitarbeiter*in individuell dabei hilft, Termine zu organisieren

oder Dokumente im Intranet zu finden. Lokale Daten sind hier Sprachaufnahmen von Mitarbeiter*innen und andere Informationen (z. B. mündliche Korrekturen der Mitarbeiter*innen, die dem Sprachassistenten einen offensichtlichen Fehler aufzeigen). In der technischen Beschreibung dieses Beispiels werden wir nicht zwischen Mitarbeiter*innen und den Geräten, auf denen ihr Sprachassistent läuft, unterscheiden.

Sprachaufnahmen sind aus verschiedenen Gründen wie zum Beispiel bei Berufsgeheimnisträgern oder im Falle von Geschäftsgeheimnissen als streng vertraulich anzusehen. In einem konservativen Verständnis sind lokale Modelle solcher Sprachaufnahmen in der Regel wie personenbezogene Daten zu behandeln, da sie Rückschlüsse auf das Gesagte zulassen. Außerdem können globale Modelle vor dem nächsten lokalen Trainieren noch personalisiert werden. Das föderierte Lernen kann in solchen Fällen globale Modelle berechnen, die gesprochenen Texte in mehreren Sprachen erkennen und in entsprechende Aktionen umsetzen, da das implizite Wissen aller lokalen Modelle des international operierenden Unternehmens im globalen Modell mit abgebildet ist. Dieses Beispiel veranschaulicht die Notwendigkeit und den Nutzen des föderierten Lernens für den Einsatz in Bereichen, die relevant für den Privatsphärenschutz sind.

6.3 Privatsphäre wahrendes föderiertes Lernen

Wir skizzieren nun eine Privatsphäre wahrende, allgemein gültige Form des föderierten Lernens, die die Techniken der Anonymisierung und Pseudonymisierung benutzt. Zur Veranschaulichung stellen wir sie aber im Kontext des obigen Beispiels 6.2 dar.

Das Beispielunternehmen beabsichtigt, föderiertes Lernen eines Sprachassistenten mit Hilfe eines externen Dienstleisters zu nutzen. Im Sinne der DSGVO ist das Unternehmen Verantwortlicher und der Dienstleister ein Auftragsverarbeiter. Vor der Entwicklung des Sprachassistenten führt das Unternehmen eine Datenschutzfolgenabschätzung durch, welche unter anderem diese Einsichten ergab:

1. Das lokale Modell einer/s beliebigen Mitarbeiter*in ist ein personenbezogenes Datum. Daher sollten weder das Unternehmen noch der Dienstleister oder andere Mitarbeiter*innen Rückschlüsse auf dieses lokale Modell oder deren Trainingsdaten ziehen können.

2. Das globale Modell soll keine Schlüsse auf die lokalen Daten oder Trainingsdaten der Mitarbeiter*innen ermöglichen.

Eine Risikobewertung zur Wahrung von Geschäftsgeheimnissen hat zusätzlich ergeben:

3. Der Dienstleister darf keine Schlüsse über die globalen Modelle, die von ihm in den Runden berechnet werden, ziehen.

Natürlich beinhaltet das Verhindern solcher Schlüsse für eine Partei, dass die entsprechenden Modelle oder Trainingsdaten dieser Partei nicht direkt zugänglich sind. Eine oberflächliche

Betrachtung dieser drei Anforderungen legt nahe, daß diese nicht umsetzbar sind. Wie kann zum Beispiel ein Dienstleister lokale Modelle zu einem globalen Modell aggregieren ohne lokale und globale Modelle zu kennen? Techniken der Anonymisierung und Pseudonymisierung können hier Abhilfe schaffen.

Für den Sprachassistenten haben lokale und globale Modelle die gleiche Struktur; sie bestehen aus vielen mathematischen Parametern und konkreten Zahlen. Wir benutzen die Variable *d* um die Anzahl dieser Parameter zu bezeichnen. Im Privatsphäre wahrenden Aggregieren werden diese Modelle in leicht umkehrbaren Schritten transformiert. Zuerst wird die topologische Struktur der Modelle in einen *d*-dimensionalen Zahlenvektor abstrahiert – zum Beispiel (1.3456, -0.456, 0.298, 4.019) für *d*=4; wobei ein Modell in der Praxis natürlich tausende oder hunderttausende solcher Parameter hat.

Anschließend werden die Zahlen in diesen Vektoren als Ganzzahlen skaliert, sodass alle Berechnungen der Aggregierung Ergebnisse liefern, die kleiner als eine geeignete Ganzzahl *m* sind. Ein/e Mitarbeiter*in mit Pseudonym *k* hat ein lokales Modell $x[k]$ welches mittels eines Zufallsvektors $r[k]$ desselben Typs pseudonymisiert wird, konkret als Summe $x[k]+r[k]$, modulo *m*. Der Vektor $r[k]$ wird im Folgenden als Maske bezeichnet. Ist zum Beispiel, stark vereinfacht, *d*=4 und *m*=99 und haben wir $x[k]$ = (45,12,7,78) und $r[k]$ = (9,63,95,23), so ergibt dies $x[k]+r[k]$ = (54,75,3,2); z. B. ist 7+95 gleich 102 = 3, modulo 99.

Diese Pseudonymisierung ist eine perfekte Verschlüsselung: Für das Ergebnis (54,75,3,2) kommt jeder mögliche Wert von $x[k]$ in Frage, da jede Differenz $x[k]$ - (9,63,95,23) ein möglicher Maskenwert $r[k]$ ist. Ist z. B. $x[k]$ = (5,73,2,59) dann ist (54,75,3,2) - (5,73,2,59) = (49,2,1,42) ein Wert der Maske r[k], sodass $x[k]+r[k]$ = (54,75,3,2) für diesen Wert von $x[k]$ gilt. Da zu jedem möglichen Wert von $x[k]$ eine solche Maske existiert die das Ergebnis (54,75,3,2) erklärt, bedeutet dies, dass man von (54,75,3,2) keinerlei Schlüsse auf das lokale Modell $x[k]$ ziehen kann ohne Hintergrundinformationen zum Wert der Maske $r[k]$ zu haben!

Der Sprachassistent der Mitarbeiter*in *k* kann daher das maskierte lokale Modell, den Vektor $x[k]+r[k]$ wie in Abbildung 7 veranschaulicht, an den Koordinator schicken. Dieser berechnet dann die Summe solcher maskierten lokalen Modelle. Bei drei Mitarbeiter*innen wäre diese Summe der Aggregation der Wert von

$(x[k1]+r[k1]) + (x[k2]+r[k2]) + (x[k3]+r[k3])$

Um den Wert des neuen globalen Modells, $x[k1]+x[k2]+x[k3]$, hieraus zu rekonstruieren, muss die Maskensumme $r[k1]+r[k2]+r[k3]$ hiervon abgezogen werden.

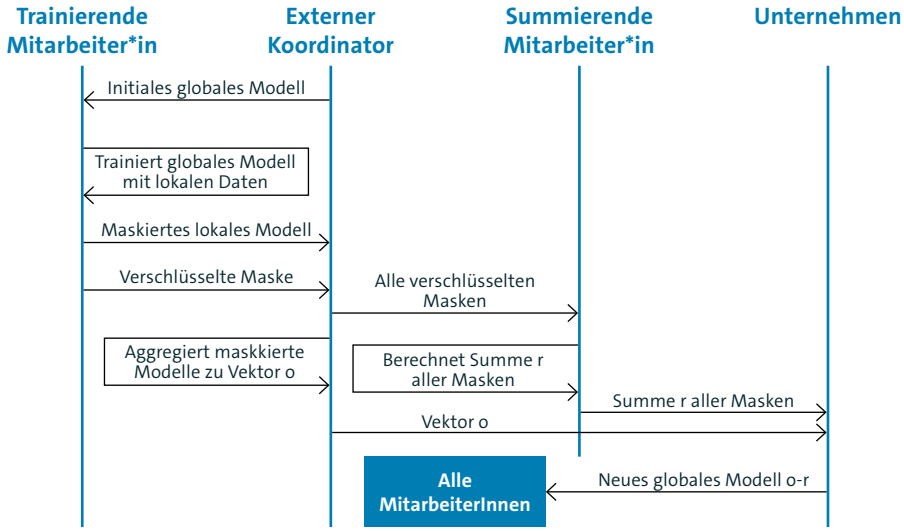

Abbildung 7: Privatsphäre wahrendes föderiertes Lernen. Mitarbeiter*innen stellen zuerst fest (nicht im Bild), ob sie trainierende oder summierende Mitarbeiter*innen sind, oder ob sie keinen Beitrag dazu leisten. Trainierende Mitarbeiter*innen berechnen aus dem globalen Modell und lokalen Daten ein neues lokales Modell und schicken ihr maskiertes lokales Modell sowie die Verschlüsselung der Maske an den Koordinator. Der Koordinator schickt die Verschlüsselungen der Masken an summierende Mitarbeiter*innen, die sie entschlüsseln, die Summe aller Masken berechnen und diese an das Unternehmen schicken. Der Koordinator schickt die Aggregierung aller maskierten lokalen Modells an das Unternehmen. Das Unternehmen berechnet aus diesen erhaltenen Vektoren das globale Modell und schickt es allen Mitarbeiter*innen zur weiteren Verwendung.

In unserem Beispiel soll der Koordinator den Wert des globalen Modells wegen der Geheimniswahrung nicht erfahren. Daher muss das Abziehen der Summe $r[k1]+r[k2]+r[k3]$ auf der Seite des Unternehmens stattfinden, wie in Abbildung 7 dargestellt. Dieses Protokolls benutzt einen kryptographischen Mechanismus. Auch wenn in diesem Kontext der Fokus nicht auf diesen Mechanismus liegt, sei erwähnt, dass er

- eine Anzahl von **trainierenden Mitarbeiter*innen** auswählt, die in der momentanen Runde ein maskiertes Modell an den Koordinator schicken dürfen.
- eine Anzahl von **summierenden Mitarbeiter*innen** auswählt, welche die Maskensumme $r[k1]+r[k2]+...$ berechnen und dem Koordinator schicken,
- wobei der Mechanismus weder vom Koordinator noch von den Mitarbeiter*innen manipulierbar ist, sondern er die jeweiligen Mitarbeiter*innen Privatsphäre wahrend und verifizierbar auswählt.

Mitarbeiter*innen sind in diesem Protokoll identifiziert durch Zufallszahlen und somit pseudonymisiert. Die Kommunikation im föderierten Lernen verläuft so, dass Mitarbeiter*innen nur Informationen vom Koordinator und dem Unternehmen erhalten und nicht *direkt* von anderen Mitarbeiter*innen. Dies erfordert eine Lösung, welche eine Maske $r[k]$ einer trainierenden Mitar-

beiter*in *k* an eine/n summierende/n Mitarbeiter*in weiterleitet, ohne dass der Koordinator Rückschlüsse über den Wert dieser Maske $r[k]$ machen kann. Ansonsten könnte der Koordinator mit dem Wissen über $r[k]$ prinzipiell Rückschlüsse über den Wert von $x[k]$ aus dem Wert von $x[k]+r[k]$ machen.

Dieses Problem können wir lösen, indem die Masken $r[k]$ mit einem semantisch sicheren Verfahren verschlüsselt und dem Koordinator geschickt werden, der diese verschlüsselten Masken den summierenden Mitarbeiter*innen weiterleitet. Der Koordinator kann hierbei keinerlei Rückschlüsse über den Wert einer Maske $r[k]$ ziehen: Semantische Sicherheit ist ein Konzept aus der Kryptographie das Folgendes gewährleistet: Die Verschlüsselung *C* einer Nachricht *M* erlaubt keinerlei Schlüsse über den Inhalt der Nachricht *M*. Davon ausgenommen sind Rückschlüsse, die man bereits allein aufgrund der Länge der Nachricht *M* machen könnte. In unserem konkreten Fall bedeutet dies, dass Rückschlüsse bereits aufgrund des Wissens darüber, wie viele Bits die Maske $r[k]$ selbst hat, möglich wären. Dies ist aber eine Information, die in diesem Beispiel jede Partei kennt und die hier keine Relevanz für die Sicherheit und den Datenschutz hat.

Der Dienstleister berechnet und schickt die Summe *o* der maskierten lokalen Modelle an das Unternehmen zurück. Die summierende Mitarbeiter*innen senden dann dem Unternehmen die entsprechende Summe *r* der Masken, damit das Unternehmen das neue globale Modell als Differenz $o - r$ berechnen und Mitarbeiter*innen zum erneuten lokalen Lernen schicken kann.

Dieses Design muss auch ein Auge haben auf Szenarien, in denen der Koordinator oder Mitarbeiter*innen das Protokoll bewusst oder unbewusst angreifen oder manipulieren. Zum Beispiel kann eine summierende Mitarbeiter*in einen falschen Wert der Maskensumme an das Unternehmen schicken. Dies kann man abwehren, indem der Koordinator mehrere summierende Mitarbeiter*innen für diese Aufgabe auswählt. So kann das Unternehmen nach einem Mehrheitsprinzip oder ähnlichem Auswahlkriterium verfahren, um aus der erhaltenen Maskensumme den korrekten Wert zu bestimmen.

Solche Erwägungen verbieten auch die Aggregierung von nur zwei maskierten Modellen. Sonst kann nämlich eine manipulierende trainierende Mitarbeiter*in $k1$ – die zu einem maskierten Modell $x[k1]+r[k1]$ beiträgt – personenbezogene Daten einer anderen trainierenden Mitarbeiter*in erfahren. Anstatt dem Koordinator eine Verschlüsselung der Maske $r[k1]$ zum Weiterleiten an summierenden Mitarbeiter*innen zu senden, schickt sie eine Verschlüsselung des Wertes $x[k1]+r[k1]$. Weder der Koordinator noch entschlüsselnde summierende Mitarbeiter*innen können diesen Angriff feststellen, da sowohl $r[k1]$ als auch $x[k1]+r[k1]$ ihnen als Zufallszahlen erscheinen. Das Abziehen der Maskensumme liefert dann aber

$$(x[k1]+r[k1]) + (x[k2]+r[k2]) - ((x[k1]+r[k1])+r[k2]) = x[k2]$$

und nicht wie erwünscht

$$(x[k1]+r[k1]) + (x[k2]+r[k2]) - (r[k1]+r[k2]) = x[k1]+x[k2]$$

Daher wird hier $x[k2]$, das als lokales Modell der Mitarbeiter*in $k2$ personenbezogene Daten enthält, als »globales« Modell allen Mitarbeiter*innen zum nächsten lokalen Lernen geschickt. Diese Möglichkeit und ihre arithmetischen Varianten sind daher unbedingt zu vermeiden. Konkret wird das dadurch verhindert, dass eine Runde wiederholt wird, wenn nur zwei maskierte lokale Modelle den Koordinator erreichen.

Weder das Unternehmen noch summierende Mitarbeiter*innen können aus der erhaltenen Summe r der Masken Rückschlüsse auf die Summanden ziehen: Die Summe von Zufallszahlen sagt nichts über die Werte der Summanden aus. Ist für m=99 diese Summe zum Beispiel 56, als Auswertung von $r[k1]+r[k2]+r[k3]$, so kann $r[k1]$ jeden Wert zwischen 0, 1, ..., 98 annehmen, und das gleiche gilt für $r[k2]$ oder $r[k3]$. Ist $r[k1]$ zum Beispiel 13, so ergeben $r[k2]$=27 und $r[k3]$=16 die erwünschte Summe 56. Bei mehr als zwei Summanden kann man im Allgemeinen überhaupt nichts über den Wert zweier Summanden schließen, auch wenn der Wert der Summe und der von n-2 ihrer Summanden bekannt ist.

Dies ist nützlich für den Fall, dass einige dieser Parteien kooperieren, um den Datenschutz des Sprachassistenten zu korrumpieren. Zum Beispiel kann man vom Wissen der Werte von r und $x[k]+r[k]$ keinerlei Rückschlüsse über $r[k]$ und daher auch keinerlei Schlüsse über die personenbezogenen Daten $x[k]$ ziehen. Wir wenden uns nun solchen Sicherheitsaspekten zu.

6.4 Sicherheitsaspekte des föderierten Lernens

Die Anonymisierungen von Privatsphäre wahrenden Verfahren des föderierten Lernens schaffen auch neue Angriffsflächen. Dies ist ein zwar bedauerlicher aber wohl unvermeidlicher Zielkonflikt: Sicherheit verlangt mehr Kontrolle und mehr Kontrolle verlangt mehr Information über die Akteure und Prozesse – was den Datenschutz negativ beeinflussen kann. Das trifft auch für das im Abschnitt 6.3 beschriebene Verfahren zu, wie wir schon für den Fall der Aggregierung von nur zwei Modellen diskutiert haben. Eine Bewertung solcher möglichen Bedrohungen sollte aber auch den vertraglichen Rahmen des konkreten föderierten Lernens einbeziehen. In unserem Beispiel würden die Mitarbeiter*innen des Unternehmens wohl einen Vertragsbruch begehen oder sich sogar strafbar machen, wenn sie versuchten, das Privatsphäre wahrende Protokoll aktiv zu manipulieren.

Diese Überlegungen zur Bewertung möglicher Angriffe schließen auch Szenarien ein, in denen Mitarbeiter*innen Geheimnisse oder personenbezogene Daten – seien es Masken, Modelle oder andere vertrauliche oder persönliche Daten – an unbefugte Dritte außerhalb der Protokollvorgaben weiterleiten. So könnte zum Beispiel eine Mitarbeiter*in rechtswidrig ihr/sein globales Modell an einen Mitbewerber schicken.

Solche Risiken werden also durch rechtliche Rahmen und Eingrenzungen der technischen Konfiguration (z. B. keine Aggregierung von weniger als drei maskierten lokalen Modellen) in Schranken gehalten. Eine holistische Risikobewertung sollte hier auch Bezug nehmen auf die konkreten Modellstrukturen und Algorithmen, die im föderierten Lernen eingesetzt werden. Je nach ver-

wendetem Algorithmus kann es zum Beispiel möglich sein, von Modellen Rückschlüsse über die Eigenschaften von Trainingsdaten zu ziehen – etwa, ob ein gewisser Dateneintrag in den Trainingsdaten vorkam. Die Pseudonymisierung der Mitarbeiter*innen und die Konfiguration des föderierten Lernens helfen, dass solche – theoretisch möglichen – Rückschlüsse nicht mehr machbar sind. Diese Pseudonymisierungen unterstützen dadurch Unternehmen bei der Umsetzung und Gewährleistung der DSGVO. Solche Bewertungen müssen daher alle Mittel, die nach Ermessen wahrscheinlich genutzt werden, berücksichtigen.

Darüber hinaus ist das Verständnis wichtig, ob solche Angriffsmöglichkeiten hauptsächlich eher die Performanz des gelernten Modells und gegebenenfalls weniger Aspekte des Datenschutzes betreffen. Bei gewissen Algorithmen, wie solchen die ein Wort zum Komplettieren eines Auslösesatzes lernen, kann ein Angreifer des föderierten Lernens zum Beispiel die Kontrolle über die Ausgaben einiger Eingaben des Modells erhalten. Ein Angreifer kann für den Auslösesatz »Mein Lieblingsbier ist …« z. B. eine Biermarke seiner Wahl erzwingen.

Nach dieser Beschreibung einer Privatsphäre wahrenden Lösung für das föderierte Lernen können wir uns nun einer rechtlichen Bewertung des Ganzen zuwenden.

6.5 Rechtliche Bewertung

Das Trainieren von KI-Modellen mit personenbezogenen Daten steht in ständigem Konflikt mit dem Datenschutz. Nicht nur das Finden der richtigen Rechtsgrundlage bereitet Probleme (a)), sondern auch der häufig anzutreffende Personenbezug von KI-Modellen (b)). Als Lösung haben sich die Anonymisierung und die Nutzung von synthetischen Daten (c)) etabliert. Föderiertes Lernen hingegen hat diese Nachteile nicht, erlaubt aber dennoch ein datenschutzkonformes Erstellen von KI-Modellen (d)).

a) Datenschutzrechtliche Grenzen beim Trainieren von Modellen

Die DSGVO verlangt für jede Verarbeitung von personenbezogenen Daten eine Rechtsgrundlage, also auch dann, wenn ein Unternehmen mit Daten seiner Mitarbeiter*innen Modelle trainiert. Sie ist in der Praxis aber nicht einfach zu finden. Eine Einwilligung nach Art. 6 (1) 1 lit. a DSGVO kommt zwar in Betracht, wird von Betroffenen aber in der Regel nicht abgegeben werden, sie ist schließlich freiwillig. Auch ist das Trainieren von Modellen nicht erforderlich für die Erfüllung des Arbeitsvertrages der Mitarbeiter*innen (Art. 6 (1) 1 lit. b DSGVO, § 26 (1) 1 BDSG), was etwa der Fall wäre bei der Verarbeitung von Kontodaten für die Gehaltsauszahlung. In der Praxis werden sich Unternehmen stattdessen häufig auf die sogenannten überwiegenden, berechtigten Interessen nach Art. 6 (1) 1 lit. f DSGVO berufen, unterstützt durch die Privilegierung, personenbezogene Daten zu statistischen Zwecken verarbeiten zu dürfen.[18]

18 Siehe ausführlich zur Rechtsgrundlage für das Trainieren von KI-Modellen: Kaulartz in Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, Kapitel 8.9 mit weiteren Nachweisen.

Eine Verarbeitung personenbezogener Daten ist nach der Interessenabwägung zulässig, wenn die Verarbeitung zur Wahrung der berechtigten Interessen des Verantwortlichen erforderlich ist, sofern nicht die Interessen oder Grundrechte und Grundfreiheiten der Mitarbeiter*innen, die den Schutz personenbezogener Daten erfordern, überwiegen. Es lässt sich umso besser argumentieren, dass diese Interessen den Interessen des Unternehmens unterliegen, je weniger Rückschlüsse aus den Modellen auf die Mitarbeiter*innen gezogen werden können. Relevant bei der Beantwortung dieser Frage ist nicht nur die Menge an verarbeiteten Daten und deren Schutz, sondern auch, wer Zugriff auf die Daten erhält und natürlich welchen Inhalt sie haben. Kurz: Der zwangsläufig einhergehende Eingriff in die Grundrechte der Mitarbeiter*innen sollte geringstmöglich ausfallen, in abhängig des mit der Verarbeitung zu erreichenden Zwecks.

Die Frage nach der für das Training notwendigen Rechtsgrundlage ist dabei stets vom konkreten Anwendungsfall abhängig, an dem auch die Interessenabwägung ausgerichtet wird. Wer KI-Modelle zu Überwachungszwecken trainiert, muss sicherlich stärker argumentieren, als wenn die Modelle dem Arbeitsschutz dienen sollen.

b) KI-Modelle als personenbezogene Daten

Beim Trainieren mit personenbezogenen Daten ergibt sich noch ein weiteres, ganz grundsätzliches Problem: “KI-Modelle können ebenfalls personenbezogen sein – auch, wenn es sich bei ihnen nur um eine Menge von Zahlen handelt. Grund ist die weite Definition des Begriffs der personenbezogenen Daten. Nach Art. 4 Nr. 1 DSGVO fallen darunter alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Es geht also nicht nur um den Namen einer Person, sondern um alles, was Rückschlüsse auf eine natürliche Person zulässt. In der KI-Forschung hat sich gezeigt, dass es durch sogenannte Model-Inversion-Attacks[19] möglich ist, von KI-Modellen Rückschlüsse auf die Trainingsdaten zu ziehen. Mit Membership-Inference-Attacks[20] kann außerdem festgestellt werden, ob konkrete personenbezogene Daten Teil der Trainingsdaten gewesen sind. Sind solche Angriffe erfolgreich, würden die KI-Modelle mit der DSGVO »infiziert«, denn sie wären personenbezogen und die Verantwortlichen müssten insoweit alle Vorschriften der DSGVO berücksichtigen, so etwa auch die Notwendigkeit einer Rechtsgrundlage oder die Betroffenenrechte, einschließlich etwaiger Löschungsansprüche.

Nicht jeder erfolgreiche Angriff führt indes gleich dazu, dass ein Modell personenbezogen ist. Um festzustellen, ob eine natürliche Person identifizierbar ist, sollen nach Erwägungsgrund 26 S. 3 der DSGVO nämlich nur jene Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren. Dies verlangt in der Praxis nach einer

19 Fredrikson/Jha/Ristenpart, Model Inversion Attacks that Exploit Confidence Information and Basic Counter-measures, 2015, ↗**https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf**; Veale/Binns/Edwards, Algorithms that remember: model inversion attacks and data protection law, 2018, ↗**http://dx.doi.org/10.1098/rsta.2018.0083**.

20 Shokri/Stronati/Congzheng/Shmatikov, ↗**https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7958568**.

eingehenden Bewertung, vielleicht sogar eingebettet in eine Datenschutzfolgenabschätzung (Privacy Impact Assessment) nach Art. 35 DSGVO. Überdies müssen die zum Schutz vor solchen Angriffen umgesetzten technischen und rechtlichen Maßnahmen ständig überprüft werden, da maßgeblich einzig der Zeitpunkt jedes einzelnen Zugriffs auf ein Modell ist, nicht nur der Zeitpunkt des Erstellens eines Modells.

c) Möglichkeiten zur Wahrung der Privatsphäre

In der Praxis haben sich drei Verfahren etabliert, um datenschutzkonformes Trainieren zu ermöglichen: Anonymisierung, synthetische Daten, föderiertes Lernen:

- Die Anonymisierung nimmt den Trainingsdaten ihren Personenbezug und macht die DSGVO damit unanwendbar, ist aber vergleichsweise aufwendig und geht mit einem hohen Informationsverlust einher, gerade bei sehr großen Datenmengen (z. B. müsste die Wohnadresse ersetzt werden durch das Wohnviertel).

- Synthetische Daten sind der Versuch, den Gehalt von personenbezogenen Trainingsdaten in fiktiven Trainingsdaten nachzubilden. Da synthetische Daten damit völlig fiktiv und damit anonym sind, fände die DSGVO auf sie ebenfalls keine Anwendung. Nicht vergessen werden darf jedoch, dass zur Erstellung der fiktiven Daten auch personenbezogene Daten verarbeitet werden müssen, was den Verantwortlichen nicht von der Notwendigkeit einer rechtlichen Grundlage befreit.

- Schließlich wird das oben betrachtete föderierte Lernen diskutiert.

d) Vorteile des föderierten Lernens für den Datenschutz

Attraktiv am föderierten Lernen ist die Tatsache, dass mit den rohen Trainingsdaten gearbeitet werden kann, ohne, dass diese vorher anonymisiert werden müssten. Das ist möglich, weil beim föderierten Lernen das Training lokal durchgeführt wird, zum Beispiel auf den Endgeräten der Mitarbeiter*innen, wo die Daten anfallen, und nur die Modelle zum Verantwortlichen übertragen werden. Der Verantwortliche kommt mit den rohen Trainingsdaten also niemals in Berührung, was innerhalb der genannten Interessenabwägung ein gewichtiges Argument zu Gunsten des Verantwortlichen ist.

Da die unter b) genannten Angriffe, die zur Identifizierung einzelner Mitarbeiter*innen durchgeführt werden könnten, ein gewisses Wissen über das in Frage stehende KI-Modell verlangen und damit mitunter recht aufwendig sind, ist es manchmal ohnehin schon so, dass diese nicht »nach allgemeinem Ermessen wahrscheinlich genutzt werden«. Die in diesem Kapitel skizzierte Privatsphäre wahrende Form des föderiertes Lernens verfügt aber noch über ergänzende Mechanismen, welche die Wahrscheinlichkeiten der unter b) genannten Angriffe drastisch reduzieren und die Modelle damit als anonym und nicht mehr personenbezogen zu qualifizieren können:

- Gemeint ist zunächst die Verwendung der unter 6.3 erläuterten Maskierungen. Da es ohne Kenntnis der Masken mathematisch unmöglich ist, aus den Maskierten Modellen Rückschlüsse auf die konkreten Modelle (und damit die Trainingsdaten) zu ziehen, sind die maskierten Modelle für den Koordinator nicht personenbezogen. Daran ändert auch die Tatsache nichts, dass die summierenden Mitarbeiter*innen die Masken kennen, denn deren Wissen ist dem Koordinator nicht zurechenbar. Es ist vielmehr bei der Frage zu berücksichtigen, ob es »nach allgemeinem Ermessen wahrscheinlich genutzt werden« wird.

- Die Identifizierung einzelner Mitarbeiter*innen wird außerdem dadurch erschwert, dass beim Trainieren mit Trainingsdaten mit Modellen gearbeitet wird, die selbst schon das aggregierte »Wissen« über zahlreicher (mindestens drei) andere Trainingsdatensätze enthalten. Da jede weitere Aggregation Rückschlüsse auf einzelne Trainingsdatensätze erschwert, trägt dies zur Anonymisierung der Modelle bei.

e) Zusammenfassung

Föderiertes Lernen bietet aus datenschutzrechtlicher Sicht große Vorteile. Im Gegensatz zum Anonymisieren von Daten kann beim föderierten Lernen mit den ungeschwärzten Rohdaten trainiert werden. Im Gegensatz zu den synthetischen Daten muss kein Datensilo geschaffen werden, sondern Modelle werden dort trainiert, wo die Daten anfallen. Der Verantwortliche kommt mit den personenbezogenen Trainingsdaten nicht in Berührung, sondern erhält nur, was ihn interessiert, nämlich die trainierten Modelle.

Diese sind nicht personenbezogen, denn beim hier dargestellten föderierten Lernen werden einige Maßnahmen hintereinandergeschaltet, die schon einzeln und für sich betrachtet geeignet sind, den Personenbezug der Modelle zu beseitigen, erst Recht aber in der Summe. Das erlaubt eine Verarbeitung der Modelle ohne Restriktionen durch die DSGVO.

7 Anonymisierung und Pseudonymisierung von Medieninhalten: Risiken und Gegenmaßnahmen

7 Anonymisierung und Pseudonymisierung von Medieninhalten: Risiken und Gegenmaßnahmen

Verena Battis, Lukas Graner, Martin Steinebach, Patrick Aichroth

Rekonstruktion von Trainingsdaten, Model Inversion, Membership Inference

In diesem Kapitel soll der Schutz von personenbezogenen Daten bei multimedialen Inhalten betrachtet werden, und hier genauer auf den in den Medien enthaltenen ableitbaren Informationen und nicht, wie ebenfalls möglich, aus den Metadaten. Wie bereits in der Einleitung erwähnt, kann allgemein für Bild- und Videomaterial eine Anonymisierung u.a. durch Vergröberung (z. B. starkes Verpixeln der Gesichtsregion oder Substitution (z. B. schwarzer Balken über dem Gesicht) erreicht werden. Für Audiodaten bzw. Sprachmaterial kann je nach Anwendungsfall die Stimme bei Beibehaltung der Sprachinhalte »anonymisiert« werden, z. B. durch eine Verfremdung der Stimme, Sprachsynthese oder Ersetzen der Sprechercharakteristik mittels Voice Conversion[21]. Es existieren aber auch Verfahren, die sich auf die linguistische Auffälligkeiten bzw. Verflachung des Vokabulars konzentrieren[22], oder die Sprache unwahrnehmbar machen oder völlig entfernen[23].

Betrachtet man die Frage von Anonymisierung und Pseudonymisierung im Kontext des maschinellen Lernens, liegt der Fokus oft auf textuellen Daten und Datenbanken, und auf kritischen Informationen wie Namen, Adressen, Gesundheitsdaten, oder IP Adressen. Das liegt daran, dass die Verarbeitung textueller Daten traditionell im Fokus der KI-Entwicklung steht. Maschinelles Lernen hat inzwischen aber auch gerade bei der Verarbeitung multimedialer Inhalte enormes Potenzial bewiesen:

Aus Medieninhalten werden mittels KI heute wertvolle Metadaten wie Inhaltsbeschreibungen, Ortsangaben, Stimmungen oder Sprachtranskriptionen abgeleitet. Möglich wird dies durch große Mengen annotierter Medien, mit denen Netze trainiert werden können.

Aus diesem Grund sind die Themen Anonymisierung und Pseudonymisierung auch für die Verarbeitung von Medieninhalten relevant, und mit besonderen Herausforderungen verbunden:

21 D. Wu and H. Lee, »One-Shot Voice Conversion by Vector Quantization,« ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7734-7738, doi: 10.1109/ICASSP40776.2020.9053854

22 G. Zhang, S. Ni and P. Zhao, »Enhancing Privacy Preservation in Speech Data Publishing,« in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.2983228

23 D. Liaqat, E. Nemati, M. Rahman and J. Kuang, »A method for preserving privacy during audio recordings by filtering speech,« 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 79-82, doi: 10.1109/LSC.2017.8268148

Bilder, Videos und Audiodaten transportieren z. B. über Gesichter, Stimme und die Sprache selbst Informationen, die mit Blick auf den Schutz der Privatheit kritisch sind. Gleichzeitig sind diese Informationen für die entsprechenden Analysen in vielen Fällen gar nicht relevant, z. B. im Fall von akustischer Maschinenüberwachung.

Wenig beachtet wurde bisher allerdings, dass auch beim Training von KI erhebliche Risiken für die Privatheit existieren, gerade im Kontext von Mediendaten. Jüngere Forschungsergebnisse zeigen, dass bei trainierten Netzen die Gefahr besteht, dass sie kritische Informationen preisgeben. Zum einen kann in bestimmten Fällen das verwendete Originalmaterial zumindest näherungsweise rekonstruiert werden [1]. Zum anderen kann aber auch geprüft werden, ob bekanntes Material zum Training verwendet wurde [2]. Ein Risiko besteht in solchen Fällen vor allem dann, wenn relevante personenbezogene Informationen mit Medieninhalten verknüpft werden können. Würde bspw. eine Klinik eine Früherkennung von Krankheiten auf Basis von Portraitfotos entwickeln und das trainierte Netz dann in einer App frei zur Verfügung stellen, so könnten Portraits rekonstruiert und die so abgeleiteten Personen mit der Krankheit verknüpft werden.

Um dieses Risiko zu mindern, gilt es Methoden zu entwickeln, die die Privatheit der Datensubjekte zuverlässig schützt, ohne dabei die Klassifizierungsergebnisse signifikant zu beeinträchtigen.

7.1 Risiken in trainierten Netzen

Im Zeitalter von Big Data und maschinellem Lernen (ML) ist es noch schwieriger geworden, Privatheit zu gewährleisten, da in großen Datenbeständen – selbst in solchen aus gering strukturierten oder gar unstrukturierten Daten – entscheidende Verknüpfungen gefunden werden können, welche das Herstellen von Personenbezügen ermöglichen.

Maschinelles Lernen ist ein Teilgebiet der künstlichen Intelligenz und beschreibt eine Reihe von Lernalgorithmen, die versuchen Strukturen in Daten zu erkennen, um basierend auf diesen Mustern bspw. Klassifizierungs- oder Regressionsaufgaben zu lösen. Der Einsatz von Verfahren des maschinellen Lernens bietet sich immer dann an, wenn die zu lösenden Probleme zu komplex oder zu umfassend sind, um sie analytisch beschreiben zu können [3]. Gleichzeitig bedeuten größere Datenmengen auch, dass mehr Informationen zum Trainieren der Lernalgorithmen zur Verfügung stehen, was tendenziell zu besseren Modellen und effizienteren Schätzungen führt [4]. Neuronale Netze finden aufgrund ihrer Flexibilität und guten Generalisierungsfähigkeit in den verschiedensten Bereichen Anwendung – ob im Verarbeiten und Analysieren natürlicher Sprachen, zur Bild- oder Gesichtserkennung oder zum Aufspüren von Anomalien.

Da ML-Algorithmen üblicherweise auf disjunkten Datensätzen trainiert und evaluiert werden, wurde lange fälschlicherweise angenommen, dass es nicht möglich ist, vom finalen Modell Rückschlüsse auf die zum Training verwendeten Daten zu ziehen, was folglich einer Anonymisierung des verwendeten Datenmaterials gleichkommen würde.

Bestimmte ML-Techniken können sich jedoch unerwartet deutlich an die zum Training des Modells verwendeten Daten erinnern. So speichern Support Vektor Maschinen oder k-nächste-Nachbarn Klassifikatoren Informationen über die zum Lernen verwendeten Daten in dem Modell selbst ab. Diese sogenannten Feature-Vektoren erlauben unter bestimmten Umständen Rückschlüsse auf die Rohdaten und stellen somit ein entscheidendes Risiko dar [5].

Aktuelle Forschungen haben ergeben, dass auch bei Neuronalen Netzen das Risiko besteht, dass eine unerwartet klare Erinnerung an die zum Training verwendeten Daten im Netz verbleibt. Diese Informationen können von Angreifern genutzt werden, um Rückschlüsse auf die Trainingsdaten zu ziehen und somit die Privatheit der Datensubjekte zu gefährden.

Im Folgenden werden drei Arten von Rückschlüssen und die korrespondierenden Angriffe vorgestellt: Model Inversion, Membership Inference sowie Model Extraction.

Zur Veranschaulichung gehen wir dabei von dem Szenario aus, dass eine Partei A ein Machine-Learning-Modell auf einem vertraulichen und nicht weiter veröffentlichten Datensatz trainiert und das Modell anschließend zur Nutzung bereitstellt. Hier muss unterschieden werden, ob das Modell vollständig veröffentlicht wird oder ob dem Nutzer lediglich Zugriff auf das Modell gewährt wird, bspw. über eine API. Wird das Modell an sich veröffentlicht, kann der Nutzer das Modell nach Belieben befragen und besitzt darüber hinaus volles Wissen über den verwendeten Algorithmus, die Architektur und die Parameter des Modells. Man spricht in diesem Kontext von einem White-Box Zugriff. Im sogenannten Black-Box Setting kann der Nutzer das Modell zwar ebenfalls mit seinen eigenen Datenpunkten befragen, um eine Ausgabe zu erhalten, verfügt aber darüber hinaus über keinerlei Wissen bezüglich des verwendeten Modells, dessen Architektur oder verwendeter Parameter.

Weiter gehen wir davon aus, dass die Ausgabe des Modells aus Wahrscheinlichkeits- bzw. Konfidenzwerten besteht. Zum einen geben diese Werte an, welcher Klasse bzw. welchem Attribut der eingegebene Datenpunkt zugeordnet wird. Zum anderen bedeuten höhere Werte auch, dass sich das Modell bezüglich seiner Entscheidung sicherer ist. Resultiert in einem Klassifizierungsproblem mit n Klassen bspw. eine Wahrscheinlichkeit von 1/n für eine positive Ausgabe, so ist sich das Modell wesentlich unsicherer bezüglich seiner Entscheidung, als wenn es einen Datenpunkt mit einer Wahrscheinlichkeit von 0,98 einer der Klassen zuweist.

7.1.1 Model Inversion

Die Idee der Model Inversion ist es, das Modell selbst zu nutzen, um gezielt Datenpunkte zu rekonstruieren, die zum Training verwendet wurden. Je nach Intention des Angreifers bedarf es nicht einmal zwangsläufig einer vollständigen Rekonstruktion der Daten, sondern nur bestimmter Eigenschaften. Eine vollständige und perfekte Rekonstruktion des Trainingsdatensatzes würde eine massive Verletzung der Privatheit der Datensubjekte darstellen.

Beispielszenario

Betrachten wir die Software *Faception*, welche von dem gleichnamigen Israelischen Konzern vermarktet wird [6]. *Faception* ist ein maschinell-lernendes Modell, welches anhand von Portraits Rückschlüsse auf die Persönlichkeit der jeweiligen Person schließt. Die Entwickler werben damit, dass ihr Modell anhand eines einfachen Fotos entscheiden kann, ob es sich hierbei um einen Wissenschaftler, einen Bingo-Spieler, einen Pädophilen oder um einen Terroristen handelt.

Gerade mit Blick auf die beiden letztgenannten Kategorien kann ein Angreifer ein besonders hohes Interesse daran haben, die zum Training des Modells verwendeten Gesichtsbilder möglichst akkurat zu rekonstruieren.

Technischer Hintergrund

Da der Angreifer bei dieser Form des Angriffs im Extremfall von einem kleindimensionalen Ergebnisvektor (i.d.R. *n* Wahrscheinlichkeitswerte gemäß der *n* zuzuordnenden Klassen) auf einen hochdimensionalen Input zurückschließen muss, steigen die Erfolgschancen des Angriffs je mehr Informationen bezüglich des Modells dem Angreifer vorliegen. Im Optimalfall verfügt der Angreifer über einen *White-Box Zugriff* und kann Gradienten im Neuronalen Netz direkt berechnen. Diese Gradienten, also alle partiellen Ableitungen einer multivariaten Funktion, symbolisieren den Effekt, den eine marginale Änderung in den Inputwerten (wie etwa einzelne Pixelfarbwerte eines Bildes) auf die Ausgabe des Modells und stellen die Grundlage vieler Model-Inversion-Ansätze dar.

Um zu obigen Beispielszenario zurückzukehren: Es ist bekannt, dass das Modell Gesichtsbilder *n* unterschiedlichen Klassen zuordnet. Ein Angreifer wäre demnach daran interessiert zu wissen, wie eine Person aussieht, die bspw. zum Training der Klasse »Terrorist« verwendet wurde.

Ein Model Inversion Angriff kann nun wie folgt durchgeführt werden: Ausgehend von einem »leeren« Start-Bild (bspw. ein vollständig schwarzes Bild) wird das Modell wiederholt befragt. Verfügt der Angreifer über einen *White-Box Zugriff* kann dieser nicht nur sämtliche Ausgabewerte beobachten, sondern auch die Gradienten berechnen. Mit Hilfe dieser Gradienteninformation modifiziert der Angreifer schrittweise die Pixelwerte des Eingabebildes dahingehend, dass die Ausgabekonfidenz für die gesuchte Klasse (hier: *Terrorist*) maximiert wird. Das so generierte Bild wird nun von dem Modell unzweifelhaft dieser Klasse zugeordnet. Hierbei wird naiverweise angenommen, dass das so generierte Bild eine Instanz der Trainingsdaten darstellt oder diese zumindest angemessen repräsentiert, was allerdings nur in seltenen Fällen korrekt ist, wie im Folgenden demonstriert.

In der akademischen Literatur stößt man häufig auf das Beispiel von Fredrikson et al., die die Trainingsdaten eines Gesichtserkennungsmodells rekonstruiert haben [1]. Allerdings stellt gerade dieser Fall eines Gesichtskennungsmodells einen Extremfall dar, da jede Ausgabeklasse des Modells eine individuelle Person repräsentiert.

Das heißt alle Trainingsdatenpunkte einer bestimmten Klasse sind nur unterschiedliche Fotos der gleichen Person. Diese stammen darüber hinaus aus einem sehr kleinen, standardisierten Datensatz, welcher ausschließlich aus Frontalaufnahmen besteht. Zudem handelt es sich bei dem von Fredrikson et al. betrachtetet Modell um eine stark vereinfachte Modell-Architektur, die so in der Praxis keine Anwendung findet.

Versucht ein Angreifer nun, wie oben beschrieben, einen Datenpunkt zu rekonstruieren, welcher vom Modell mit einer hohen Konfidenz der Zielklasse zugeordnet wird, dann wird hierbei im Grunde nur ein Mittelwert aller Trainingsbilder dieser Klasse gebildet und nicht wie gewollt, ein tatsächlicher Trainingsdatenpunkt rekonstruiert. Weil alle Trainingsdaten innerhalb einer Zielklasse die gleiche Person abbilden und aufgrund des hohen Standardisierungsgrades, ist das künstlich generierte Bild dennoch für Menschen wiedererkennbar (vgl. Abbildung 9, mittlere Spalte).

Sobald die Daten innerhalb der Klassen eine größere Variation aufweisen oder eine fortgeschrittenere Modell-Architektur genutzt wird (wie etwa ein neuronales Netz mit einer sehr großen Tiefe), sind die mittels Model Inversion gewonnenen Rekonstruktionen oftmals nicht mehr als Objekte, geschweige denn als spezifische Instanzen des Trainingsets wiederzuerkennen (vgl. Abbildung 8, sowie Abbildung 9 Spalte 3).

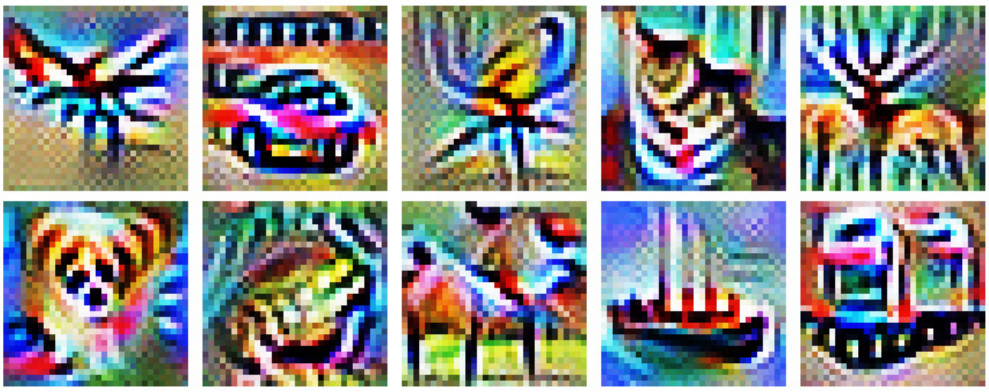

Abbildung 8: Model Inversion Angriff auf den CIFAR 10 Datensatz, in Anlehnung an [2]. Die rekonstruierten Klassen sind: Oben – Flugzeug, Auto, Vogel, Katze, Hirsch. Unten – Hund, Frosch, Pferd, Schiff, LKW.

Ergebnisse, sowohl aus aktueller akademischer, wie auch aus unserer eigener Forschung in der Abteilung Media Security und IT Forensics des Fraunhofer SIT, zeigen jedoch, dass trotz erschwerter Bedingungen – wie z. B. durch eine hohe Modellkomplexität – erweiterte Angriffsansätze dennoch Erfolge erzielen können.

Mittels dem von uns entwickelten Ansatzes »GenMoIn« lassen sich auch aus komplexeren Modellen Rekonstruktionen erzielen. Wenngleich nicht vollständig wahrheitsgetreu, können

diese für den Angreifer dennoch relevante Informationen enthalten – vor allem im Vergleich zu den stark verrauschten Bildern des Referenzangriffs (vgl. Abbildung 9, Spalte 4 und 3). So können selbst in einem nur teilweise rekonstruierten Gesichtsbild trotz fehlender oder inkorrekter Gesichtszüge, andere, potenziell sensible Informationen, wie Hautfarbe oder das Tragen einer Brille, ermittelt werden. Dieser Zusammenhang ist in Abbildung 9 dargestellt. In den Rekonstruktionen nach [1] (Spalte 3) lassen sich zwar Gesichtskonturen erkennen, diese sind jedoch sehr verrauscht und geben keine Auskunft über Details. Im GenMoIn-Ansatz (Spalte 4) wurde dagegen das Wissen, dass es sich um Portraitfotos handelt, direkt in den Rekonstruktionsprozess integriert, sodass automatisch ein erkennbares Gesicht generiert wird. Obwohl es sich nicht um exakte Replikationen von Trainingsinstanzen handelt, sind Details, wie die Haarfarbe, prägnante Gesichtszüge oder das Tragen von Accessoires deutlich erkennbar. So sind die charakteristische Nase und dunklen Haare von Person A, wie auch die volle Unterlippe und das Tragen einer Brille von Person B deutlich in den Rekonstruktionen von GenMoIn zu erkennen.

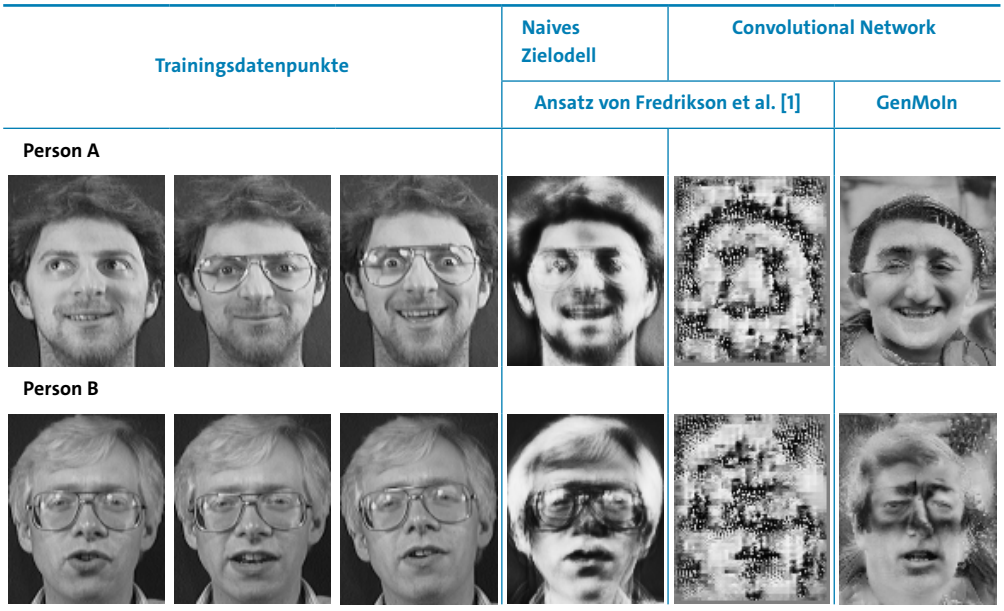

Abbildung 9: Model Inversion Angriff auf den ATT Faces Datensatz. Jede Zeile, und somit Person,repräsentiert eine individuelle Klasse. Die linke Spalte stellt jeweils eine Teilmengen der entsprechenden Trainingsbilder dar. Die zweite Spalte zeigt Ergebnisse des Angriffs von Fredrikson et al. [1]. Bei dem Zielmodell handelt es sich hierbei um eine naive Architektur, welche in der Bildverarbeitung so keine Anwendung findet. Die beiden letzten Spalten beziehen sich auf ein Zielmodell mit einer fortschrittlicheren und komplexeren Architektur, wobei links Ergebnisse des Ansatzes nach Fredrikson et al. [1] und rechts die eines unserer eigenen Angriffsansätze (GenMoIn) dargestellt sind.

Alle bisher betrachteten Beispiele und Szenarien gingen von einem *White-Box Zugriff* auf das Modell und somit von voller Einsicht des Angreifers auf die Modellparameter aus. Ist diese nicht gegeben (*Black-Box Setting*), so erweist sich die Model Inversion zwar als deutlich schwieriger, aber nicht als unmöglich. Gradienten können etwa über mehrfaches Hochladen von leicht abgewandelten Datenpunkten und Analyse der Modellausgaben geschätzt werden, was allerdings zu einem erheblichen Anstieg in den Modellanfragen und somit auch in den monetären

Kosten führen würde. Viele Dienste beschränken zudem die Nutzung (siehe dazu ↗Kapitel 2), weshalb ein solches Vorgehen ebenfalls nicht praktikabel wäre.

7.1.2 Membership Inference

Das Ziel des Membership Inference Angriffs ist es, anhand eines bestimmten Datenpunktes, eine Aussage darüber treffen zu können, ob eben jener Datenpunkt zum Trainieren des betrachteten Modells verwendet wurde. Die zugrundeliegende Aufgabe des Zielmodells, also ob es sich bspw. um eine Klassifikation oder Regression handelt, ist hierbei für den Erfolg des Angriffs unwesentlich. Einem Angreifer geht es rein um das Verknüpfen eines Datenpunktes mit zusätzlich vorhandenen Informationen über den Trainingsdatensatz. Shokri et al. [2] bewiesen, dass neuronale Netze aufgrund ihrer Konstruktion anfällig für Membership Inference Angriffe sind. Die Autoren wiesen nach, dass ein trainiertes Netz oftmals spürbar anders auf Informationen reagiert, welche bereits zum Training verwendet wurden, als auf bisher ungesehene Testdaten. Anhand dieser Rückmeldung kann ein Angreifer zuordnen, ob ein Individuum in einem bestimmten Datensatz enthalten ist oder nicht.

Allgemein stellen Angriffe wie die Membership Inference eine Verletzung der Privatheit dar, sind aber besonders dann kritisch, wenn es sich um sensible Informationen handelt, wie bspw. die finanzielle Situation oder medizinische Angaben über eine Person.

Beispielszenario

Ein Krankenhaus entwickelt ein auf maschinellen Lernverfahren basierendes Modell, welches Erbkrankheiten anhand von Portraitfotos der Patienten erkennt. Das Krankenhaus stellt das Modell nach abgeschlossenem Training der Öffentlichkeit frei zur Verfügung und betont dabei, dass es sich um eine komplett hauseigene Implementierung handele – das Modell wurde ausschließlich auf Daten trainiert, die in diesem Krankenhaus erhoben wurden.

Ein Angreifer, der im Besitz eines Portraitfotos ist, kann nun mittels Membership Inference herausfinden, ob die entsprechende Person im besagten Krankenhaus behandelt worden ist. Die Ausgabe eines solchen Angriffs ist binär – positiv, falls der Datenpunkt zum Training verwendet wurde und negativ, wenn nicht. Ein positives Ergebnis gilt allgemein als hinreichende Bedingung für die zu ermittelnde Information.

Technischer Hintergrund

Wie kommt es nun dazu, dass ein trainiertes Netz merkbar anders auf Informationen reagiert, welche bereits zum Training verwendet wurden als auf bisher ungesehene Testdaten? Das Training eines Neuronalen Netzes ist kein einmaliger, sondern ein iterativer Vorgang während dem das Modell den (endlichen) Trainingsdatensatz in unterschiedlichen Konstellationen immer wieder neu bewerten muss.

Häufig ist das Ziel des Trainings eine möglichst gute Anpassung zwischen den Modellentscheidungen und den tatsächlich beobachteten Realisationen zu erreichen. Dafür wird am Ende jeder Trainingsiteration eine Verlustfunktion (engl. Loss) berechnet, die die Abweichung zwischen geschätzten und tatsächlichen Werten misst. Während des Trainings werden die Parameter des Modells so verändert, dass die resultierende Verlustfunktion minimiert wird.

Genau hier liegt allerdings ein zentrales Problem des überwachten Lernens begraben. Wird ein Modell zu lange auf einen endlichen Datensatz trainiert, beginnt es irgendwann damit sich Trainingsdatenpunkte zu merken, um die Verlustfunktion weiter zu minimieren. Anstatt Zusammenhänge in den Daten zu erkennen, lernt das Modell die Trainingsdaten und die zugehörigen Ausgaben zu replizieren, was zu einer sinkenden Generalisierungsfähigkeit auf bisher ungesehene Daten führt. Man spricht hier von Overfitting – einer Überanpassung des Modells an die gegebenen Daten.

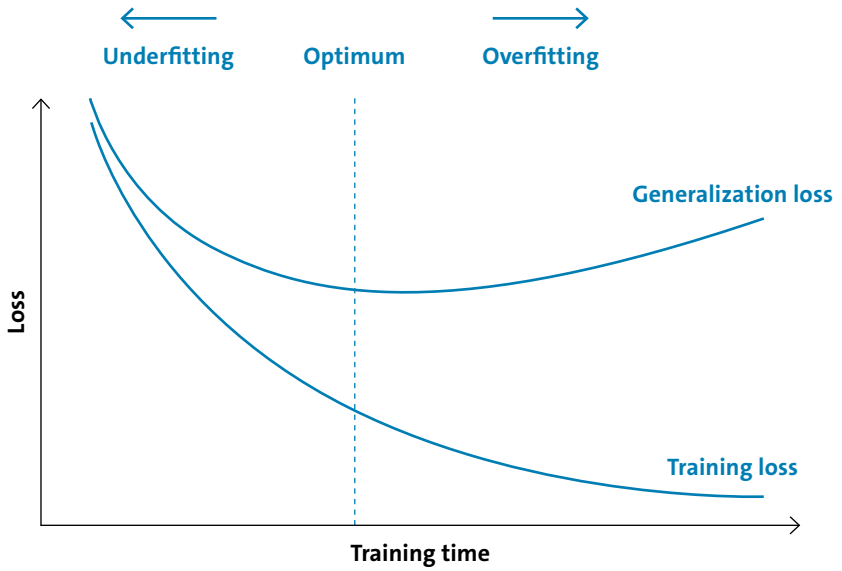

Abbildung 10: Zusammenhang Training loss und Generalisierungsfähigkeit, in Anlehnung an: Zhang, Lipton, Li, Smola, Dive into Deep Learning, 2019 [7].

Genau dieses Overfitting ist es, was sich der Membership Inference Angriff besonders zu Nutze machen kann. Sobald das Modell den optimalen Punkt überschritten hat und eine Überanpassung anfängt, beginnt es auch damit sich Trainingsinstanzen zu merken. Wird das Modell nach abgeschlossenem Training dann wiederum mit einem Trainingsdatum konfrontiert, wird es die Trainingsinstanz – vereinfacht gesprochen – wiedererkennen und mit einer höheren Konfidenz der jeweiligen Klasse zuordnen, verglichen mit einem bisher ungesehenen Datenpunkt.

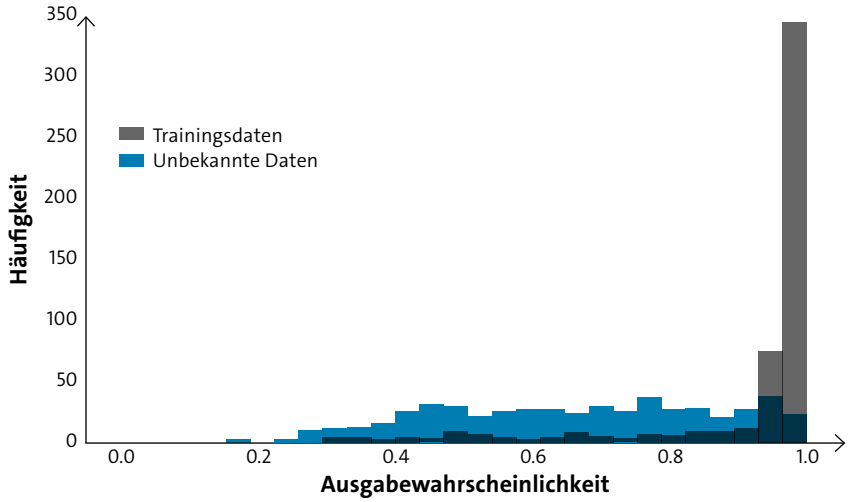

Abbildung 11: Verteilungen der Ausgabewahrscheinlichkeiten nach Trainings- und unbekannten Referenzdaten (jeweils 500 Datenpunkte) auf Purchase 10 Datensatz. Eigene Erstellung.

Graphisch dargestellt ist dieser Effekt in Abbildung 11. Es ist offensichtlich, dass die Ausgabewahrscheinlichkeiten für Elemente des Trainingsdatensatzes (grau) im Durchschnitt deutlich höher sind als die der Elemente des Referenzdatensatzes (blau) .

Der gesamte Membership Inference Angriff kann nun wie folgt ablaufen: Zunächst befragt der Angreifer das Zielmodell wiederholt, um einen vollständigen Datensatz mit Eingabe und zugehöriger Ausgabe zu generieren. Anschließend wird ein sogenanntes *Schattenmodell*, welches das Zielmodell in seiner Funktionalität bestmöglich approximieren soll, auf einer Teilmenge eben dieses Datensatzes trainiert. Für ein weiteres *Angriffsmodell*, welches für die eigentliche Erkennung der Trainingszugehörigkeit zuständig ist, muss nun zunächst noch das *Schattenmodell* mit den eigenen Trainingsdaten und mit einem bisher ungesehenen Referenzdatensatz befragt werden. Die Ausgaben des Schattenmodells bezüglich dieser beiden Datensätze, zusammen mit einem binären Code, ob es sich bei dem jeweiligen Datenpunkt um eine Trainingsinstanz handelt oder nicht, dient dem *Angriffsmodell* als Trainingsdatensatz.

Auf obiges Beispielszenario angewendet, würde der Angreifer das Modell des Krankenhauses mit einem Portraitfoto befragen und daraufhin einen Vektor an Wahrscheinlichkeiten bezüglich der einzelnen Erbkrankheiten als Ausgabe bekommen. Diesen Ausgabevektor muss er nur noch in das fertig trainierte Angriffsmodell eingeben. Im Bezug auf Abbildung 11, könnte das Angriffsmodell etwa einen einfachen Schwellenwert erlernt haben und ausgeben, dass ein Datenpunkt Teil des Trainings war, wenn die maximale Ausgabewahrscheinlichkeit des Ausgabevektors über 0,6 liegt, wobei sich der Angreifer umso sicherer sein kann, je weiter sich der Wert vom Schwellenwert entfernt.

7.1.3 Model Extraction

Allgemein ist das Ziel eines Model Extraction Angriffs das Verhalten und somit die prädiktive Leistung eines Zielmodells auf einen bisher unbekannten Datensatz zu approximieren bzw. im günstigsten Fall zu kopieren. Alternativ kann es für den Angreifer auch Sinn machen, die Architektur des Zielmodells zu *stehlen*. Aber warum denn ein Modell oder dessen Aufbau stehlen, wenn es doch frei verfügbar bzw. nutzbar ist? Neuronale Netze können sehr komplexe Funktionen approximieren und auch Zusammenhänge in großen Datenmengen finden, die ansonsten vermutlich unbekannt geblieben wären. Doch das Entwickeln sowie das Training eines solchen Netzes kann sehr zeit- und ressourcenintensiv sein. Zudem bedarf es einiges an tiefergehenden Verständnisses darüber, wie Neuronale Netze Informationen verarbeiten. Nicht jeder, der solche Techniken anwenden möchte, verfügt über die nötige Rechenleistung, das Fachwissen oder die Menge an – potenziell sensiblen – Daten, die benötigt werden, um ein ML Modell zuverlässig trainieren zu können.

Wir identifizieren demnach drei unterschiedliche Motivationen, warum ein Angreifer ein Modell *stehlen* möchte:

- Machine Learning as a Service (MLaaS) Anbieter wie bspw. Google, Amazon oder Microsoft Azure ermöglichen via API-Zugang Zugriff auf hochperformante Modelle, die gegen eine geringe Nutzungsgebühr auf die eigenen Daten angewendet werden können. Je nachdem wie viele Datenpunkte der Angreifer vom Modell verarbeitet haben möchte, kann es für den Angreifer einen monetären Vorteil darstellen, das Modell zu stehlen, sodass er es unendlich oft befragen kann, ohne weitere Gebühren für den Dienst zahlen zu müssen.

- Gleichsam kann ein besonders gut performendes Modell auch einen Wettbewerbsvorteil darstellen, sodass ein Konkurrent ebenfalls Interesse an dessen Aufbau und Wirkungsweise hat.

- Die zuvor beschriebenen Angriffe – Model Inversion und Membership Inference – basieren beide darauf, dass der Angreifer entweder Zugriff auf das Modell hat, und z. B. die Gradienteninformationen unmittelbar abgreifen kann, oder zumindest über eine hinreichend gute Approximation des Zielmodells verfügt. Dementsprechend kann Model Extraction als Vorstufe für die beiden Angriffe genutzt werden, um so deren jeweiligen Erfolgschancen zu erhöhen.

Beispielszenario

Ein weiterer Angriff auf Neuronale Netze, welcher hier nicht weiter betrachtet wurde, weil er keine direkte Bedrohung für die Privatheit darstellt, sind *Adversarial Examples*. Adversarial Examples sind kleine, für den Menschen in der Regel nicht wahrnehmbare Veränderungen in den Daten, die dazu führen, dass das Modell eine andere als die erwünschte Entscheidung trifft.

Anwendungen des autonomen Fahrens verwenden fast ausschließlich Neuronale Netze, die basierend auf den gelieferten Eingaben – Video-, Sensor-, Radardaten – das zu lösende Problem

in kleinere Klassifikationsaufgaben aufspalten. Nimmt eine Kamera bspw. ein Stopp-Schild auf, liefert das Neuronale Netz die Ausgabe »Anhalten!«.

Ein böswilliger Angreifer, dem es gelingt dieses Netz hinreichend mittels Model Extraction zu approximieren, ist nun in der Lage, Adversarial Examples basierend auf den Gradienteninformationen des gestohlenen Netzes zu kreieren, die dazu führen, dass bspw. ein Stopp-Schild als ein Vorfahrtsschild missinterpretiert wird [8]. Wird dieses Adversarial Example im öffentlichen Raum angebracht, wo es von anderen selbstfahrenden Vehikeln, die das gleiche Modell wie das Zielmodell nutzen, erfasst werden kann, kann dies verheerende Konsequenzen haben.

Technischer Hintergrund

Aktuell werden in der Literatur verschiedene Ansätze verfolgt, wie die Rekonstruktion eines unbekannten Modells mit möglichst wenig Vorinformationen am erfolgreichsten erfolgen kann. Auch ist nicht immer eine vollständige Extraktion des Zielmodells samt Architektur und verwendeten Hyperparametern das Ziel. Häufig beschränken sich die Angriffe auch nur darauf einzelne Komponenten zu ermitteln – bspw. ob eine Convolutional-Layer verwendet wurde oder welche Art der Regularisierung [9].

Am häufigsten sind sogenannte *Seitenkanalangriffe* (engl. side-channel attack) [10][11][12] sowie eine Form der *Knowledge-Distillation* zu finden. Bei letzterem wird versucht das Wissen eines Modells durch wiederholtes Befragen in ein separates, meist kleineres Modell zu überführen. Häufig werden dazu adaptive Verfahren (sog. *learning strategies*) verwendet. Diese identifizieren Trainingsdatenpunkte mit möglichst hohem Informationsgehalt, sodass die Anzahl der Anfragen, die an das Zielmodell gerichtet werden müssen, um ein aussagekräftiges Modell zu destillieren, möglichst klein gehalten werden kann [13][14][15][16]. Letzteres ist aus mehreren Gründen relevant für den Angreifer. Zum einen werden durch weniger Anfragen geringere Kosten in einem Pay-per-Use System erzeugt, zum anderen ist die Gefahr, dass der Angriff als solcher identifiziert wird geringer, je weniger Anfragen an das Zielmodell gestellt werden und je weniger Zeit der Angriff an sich benötigt.

In diesem Kontext ist auch die Arbeit von Oh et al. [17] zu nennen, die einen Meta-Klassifikator trainiert haben, welcher anhand der Ausgabe des Zielmodells die verwendeten Hyperparamter, wie z. B. die Tiefe des Netzes, bestimmen sollte. In einer Erweiterung ihres Ansatzes gibt das Zielmodell durch Befragung mit einem speziell konstruierten Adversarial Example sogar selbst Informationen über die Modellspezifikationen preis.

7.2 Gegenmaßnahmen

Basierend auf dem Verständnis der im vorhergehenden Abschnitt beschriebenen Risiken ist es möglich, Schutzmechanismen gegen diese zu entwerfen. Ansätze des privacy-preserving Zugriffs auf Datenmengen (und als solche kann auch ein trainiertes Netz gesehen werden) sind gegebenenfalls auch geeignet, mit maschinellem Lernen genutzt zu werden. Dabei sind sowohl die

Trainings- als auch die Abfrage-/Entscheidungsphase mögliche Stellen, an denen die neuen Mechanismen eingebracht werden können. Sie sind also sowohl für White-Box als auch Black-Box-Szenarien anwendbar. Ziel ist es, dass trainierte Modelle keine oder nur akzeptabel geringe Informationen über die Trainingsdaten preisgeben können.

Im Folgenden wird eine Reihe solcher Mechanismen vorgestellt. Es gilt allerdings zu beachten, dass i.d.R. alle Maßnahmen zum Schutz der Privatheit mit einer Reduktion der Prognosequalität des Modells einhergehen. Bei allen Gegenmaßnahmen gilt es folglich den klassischen Trade-Off zwischen Datennutzen und Datensicherheit genauestens abzuwägen.

7.2.1 Restriktion des Outputs

Wird ein Modell ohne Einsicht in die enthaltenen Parameter zur Verfügung gestellt (Black-Box), so werden bei einer Anfrage eines Nutzers in der Regel Konfidenzwerte zu den jeweiligen Klassen zurückgegeben, also ein Ausgabevektor. Alle drei vorgestellten Angriffe beruhen dabei oft auf allen darin enthaltenen Werten, sowie auf eine optimale Genauigkeit, sodass auch hintere Nachkommastellen von Bedeutung sein können. Um sich vor solchen Angriffen besser zu schützen, ist es also möglich die Ausgabe auf unterschiedliche Weise zu beschränken. Zum einen lässt sich der Ausgabevektor nur teilweise zurückgeben, indem etwa nur die höchsten X Konfidenzwerte übrigbleiben und der Angreifer so keine Informationen über die restlichen Werte erhält. Zum anderen können auch die Werte selbst modifiziert werden. Bspw. können sie bis zu einem maximalen Grad zufällig verrauscht werden, oder auf eine bestimmte Nachkommastelle gerundet werden. Hierbei existieren inzwischen zahlreiche weitere Ansätze, die schlussendlich den Informationsgehalt des Ausgabevektors um einen bestimmten Faktor verringern.

7.2.2 Adversarial Regularization

Bei einem gewöhnlichen Trainingsprozess eines Neuronalen Netzes, werden die initialisierten Parameter schrittweise so angepasst, dass sich die Fehler der Vorhersagen für die Trainingsdaten verringern. Bei der Adversarial Regularization handelt es sich im Groben um einen erweiterten Trainingsprozess, der zusätzlich zum ursprünglichen Netz ein weiteres »Angriffs«-Netz einführt. In jeder Trainingsiteration greift das Angreifernetz das Zielnetz, bspw. Mittels Memberschip Inference, an und versucht seinen Angriffserfolg zu maximieren. Das Ursprungsnetz wird nun nicht mehr nur mit Hinsicht auf die Vorhersagefehler der Trainingsdaten angepasst, sondern auch so, dass sich die Erfolgschancen des Angreifernetzes verringern. Dies bringt eine implizite Regularisierung mit sich und führt dazu, es dem Angreifer nicht möglich ist ein besseres Angriffsmodell zu entwickeln als jenes, welches bereits vom Zielmodell während des Trainings antizipiert wurde. Es zeigt sich jedoch oft, dass das Einsetzen von Adversarial Regularization eine verringerte Genauigkeit des Zielmodells mit sich bringt, wobei das Ausmaß der Verringerung von der Komplexität des Modells und der Datengrundlage abhängt.

7.2.3 Distillation

Distillation wurde ursprünglich von Hinton et al. [18] vorgestellt. Die Idee ist, wie bereits erwähnt (vgl. ↗Kapitel 7.1.3) und wie der Name andeutet, die Essenz des erlernten Wissens eines Modells in ein separates Modell zu überführen. Dieses Vorgehen kommt einer Komprimierung gleich und kann sogar auch – vor allem wenn ein Ensemble an Modellen gegeben ist – genutzt werden, um die Genauigkeit zu erhöhen. Inzwischen hat sich gezeigt, dass der Prozess der Distillation auch genutzt werden kann, um höhere Standards in Bezug auf Datenschutz zu erzielen und somit auch die Erfolgschancen von Model Inversion und Membership Inference zu verringern.

Dabei wird das bereits trainierte Modell als »Lehrer« angesehen, von dem ein neues – das destillierte – Student-Modell lernt. Dazu wird eine nicht zwangsweise annotierte weitere Datenmenge, meist disjunkt von der ursprünglichen Trainingsmenge, durch den Lehrer klassifiziert. Der Student trainiert nun auf dieser Datenmenge und erkennt dabei die resultierenden Ausgabe(Konfidenz-)vektoren des Lehrers als Ground-Truth an. Einfach ausgedrückt approximiert der Student den Lehrer, wobei er während des Trainingsprozesses nicht in Kontakt mit den ursprünglichen Trainingsdaten kommt. Distillation stellt somit ein datenschutzkonformes Instrument dar, da der Student somit auch keinen Zugriff auf sensible Informationen aus dem Trainingsset hat.

7.2.4 Differential Privacy

Differential Privacy ist keine fixe Methode, sondern vielmehr eine Eigenschaft, welche verlangt, dass es für eine beliebige Analyse irrelevant ist, ob ein bestimmtes Datensubjekt im Datensatz enthalten ist oder nicht – in beiden Fällen sollten sich die Output-Verteilungen nicht signifikant unterscheiden. Intuitiv bedeutet dies, dass die Menge an Informationen, die über ein bestimmtes Individuum herausgefunden werden kann, limitiert wird. Differential Privacy wird in der Regel durch das Hinzufügen von Rauschen erreicht. Der Grad der Perturbation hängt von der Stärke des Einflusses des einzelnen Eintrags auf den Datensatz ab [19].

Eine mögliche Ausprägung ist die ε-Differential Privacy. Der Parameter ε kontrolliert in diesem Fall, wie groß der maximale Effekt eines Individuums auf das Ergebnis einer Analyse ist. Im Umkehrschluss quantifiziert ε aber auch, inwieweit die Privatheit eines Individuums durch die Analyse kompromittiert werden kann. Kleinere ε-Werte gehen daher mit einem höheren Grad an Privatheit, aber auch mit einer stärkeren Perturbation einher, was sich wiederum negativ auf die Qualität der Daten auswirkt [20]. Erschwerend kommt hinzu, dass keine feste Richtlinie existiert, wie der Parameter ε optimal gewählt werden soll. Die Wahl reicht von ε=0,01 bis ε=1 in akademischer Forschung bis hin zu Werten zwischen 1 und 10 in industriellen Anwendungsfällen (Google, Apple, US Census Bureau) [21].

7.2.5 Kryptographie

Allgemein handelt es sich bei Kryptographie um eine Wissenschaft, mit deren Methoden Informationen für Unbefugte unkenntlich gemacht werden. Das heißt, nur die Person – oder der

Server – für die die Information bestimmt ist, kann die Daten lesen und verarbeiten [22]. Als Kernziel der Kryptographie gilt es, die Integrität, Authentizität, Vertraulichkeit und Nichtabstreitbarkeit der Daten zu gewährleisten. Die Mindestanforderung an ein datenverarbeitendes System sollte sein, dass die Informationen zumindest während ihrer Übertragung und an ihrem Speicherort verschlüsselt werden. Ideal wäre eine Verschlüsselung auch während der Verarbeitung – die sogenannte homomorphe Verschlüsselung. Verschlüsselung bezeichnet in diesem Kontext ein Verfahren bzw. einen Algorithmus, der dazu dient Informationen unkenntlich zu machen. Jedes Verschlüsselungsverfahren benötigt mindestens einen Klartext und einen Schlüssel, um diesen Klartext zu einem Geheimtext, ein sogenanntes Chiffrat (Ciphertext), zu verschlüsseln. Für das Entschlüsseln des Geheimtexts wird mindestens ein Chiffrat und ein Schlüssel benötigt, um wieder den zugehörigen Klartext aus dem Chiffrat zu erzeugen. Der Schlüssel ist dabei lediglich eine Information, die zur Verschlüsselung bzw. Entschlüsselung verwendet wird.

Die homomorphe Verschlüsselung erlaubt, im Gegensatz zu herkömmlichen Verschlüsselungsmethoden, Rechenoperationen direkt auf den verschlüsselten Daten auszuführen, ohne diese zuvor in Klartext überführen zu müssen und sie dadurch angreifbar zu machen. Jede Operation liefert ein ebenfalls verschlüsseltes Ergebnis, das dechiffriert demjenigen entspricht, welches resultieren würde, wäre die Operation auf dem entsprechenden Klartext durchgeführt worden. Homomorphe Verschlüsselungen generieren jedoch einen signifikanten Rechenmehraufwand [23][24], der voll-homomorphe Verschlüsselung für rechenintensive Anwendungen wie maschinelles Lernen bisher unbrauchbar macht. Erste praktikable Ansätze verwenden daher die Vereinfachungen von »somewhat« homomorpher Verschlüsselung. So wenden Dowlin et al. [23] ein auf unverschlüsselten Rohdaten trainiertes Neuronales Netz auf homomorph verschlüsselten Daten an. Long et al. [24] haben das Training verschiedener ML-Verfahren mit additiv-homomorpher Verschlüsselung und Zero-Knowledge-Beweisen realisiert.

7.2.6 Sichere Mehrparteienberechnung

Sichere Mehrparteienberechnung (engl. secure multi-party computation, MPC) ist ein Teilgebiet der Kryptographie und erlaubt das gemeinschaftliche Berechnen einer Funktion, für die mehrere Parteien eine Eingabe liefern. Die Privatheit wird in dieser Art der Berechnung dadurch gewahrt, dass jede der beteiligten Parteien nur das Endergebnis, d. h. die Funktionsausgabe, und die eigene Eingabe erfährt. Die Eingaben der übrigen Teilnehmer bleiben verborgen. Je nach Anzahl der Teilnehmer, deren Dropout Wahrscheinlichkeit und tolerierbaren Kommunikationsaufwand existieren verschiedene Ansätze, um dieses Ziel zu erreichen. So kann die Berechnung der Funktion bspw. auf mehrere nicht-kolludierende Server aufgeteilt werden. In Situationen, in denen nur ein, möglicherweise nicht vertrauenswürdiger, Server zur Verfügung steht, finden häufig Verfahren der homomorphen Verschlüsselung Anwendung [25]. Die Mehrparteienberechnung findet sich häufig in Kombination mit Cloud Computing.

7.2.7 Föderiertes Maschinelles Lernen

Hitaj et al. [26] haben nachgewiesen, dass es selbst in einem vermeintlich sicheren MPC-Protokoll möglich ist, mit Hilfe eines Generative Adversarial Networks (GAN), sensible Daten über die übrigen aufrichtigen Teilnehmer zu sammeln. Vor diesem Hintergrund mag es ratsam erscheinen, Daten aus dem eigenen Kontrollbereich gar nicht erst herauszugeben, um Privatheit effektiv schützen zu können.

Eine Lösung hierfür stellt das dezentrale bzw. föderierte Lernen dar. Hierbei trainieren die Nutzer ein Grundmodell lokal auf ihren individuellen Daten und übermitteln lediglich die neu berechneten Gradienten an den Serviceprovider. In einem periodischen Prozess aktualisiert der Provider das Gesamtmodell anhand der übermittelten Informationen aller Teilnehmer und stellt es ihnen anschließend zum Download zur Verfügung. Diese trainieren nun das aktualisierte Modell erneut lokal und senden die resultierenden Gradienten zurück an den Server [27].

Allerdings ist es selbst in diesem dezentralisierten Lernansatz möglich, aus den übermittelten Beiträgen der einzelnen Teilnehmer private Informationen zu extrahieren [28]. Um diesen Anteil möglichst gering zu halten, erlaubt der Ansatz nach Shokri und Shmatikov [29], dass nicht alle Parameterupdates mit dem Server geteilt werden müssen, sondern nur eine kleine Teilmenge, deren Größe vom Nutzer selbst festgelegt wird. Zusätzlich wird in diesem Ansatz das Konzept von Differential Privacy umgesetzt. Allerdings sollte sich der Nutzer des Trade-offs zwischen Menge der geteilten Parameterupdates sowie Trainingszeit bzw. -qualität bewusst sein.

7.2.8 Datensynthese

Eine andere Richtung der Privatheit-erhaltenden Datenveröffentlichung und -analyse verfolgt der Ansatz der Differentially Private Data Synthesis (DIPS). Hierbei werden Daten auf Basis realer Datensätze bspw. mittels Copula-Funktionen [30] oder Generative Adversarial Networks [31] unter Einhaltung von Differential Privacy synthetisiert. Der offensichtliche Vorteil dieses Ansatzes ist, dass die simulierten Daten bereits Differential Privacy erfüllen und somit keine Rückschlüsse auf die Ursprungsdaten ermöglichen – im Gegensatz zu traditionellen Datensyntheseverfahren. Darüber hinaus besitzen die Daten annähernd die gleichen Verteilungseigenschaften, wie die zugrundeliegenden Originaldaten und können in beliebiger Anzahl generiert werden, um so bspw. die Güte eines ML-Models zu verbessern [32][21].

7.3 Diskussion

Angriffspunkte für die weitere Forschung allgemein sind zum einen die Anwendbarkeit der Verfahren bzw. deren mangelnde Flexibilität. Die meisten Privatheit-erhaltenden Verfahren sind nur für die Anwendung auf einen bestimmten Lernalgorithmus optimiert und auf andere ML-Verfahren nur schwer bis gar nicht anwendbar. Auf der anderen Seite stellt mangelnde Skalierbarkeit ebenfalls ein Hindernis für die Anwendung Privatheit-erhaltender Maßnahmen in der Praxis dar. Das Schützen sensibler Informationen generiert immer zusätzliche Kosten – ent-

weder aufgrund von höherem Berechnungsaufwand, extrem langen Trainingszeiten oder weil der Nutzen der Daten bspw. durch zugefügtes Rauschen vermindert wird. In manchen Fällen fallen diese Kosten sogar so hoch aus, dass eine Anwendung in der Praxis nicht tragbar ist [3].

Es ist anzunehmen, dass verschiedene Datentypen unterschiedlich stark anfällig für unterschiedliche Angriffe auf trainierte Netze sind. Multimedia-Daten sind hier potenziell besonders geeignet, da bei ihnen eine Annäherung an den Originalzustand ausreicht, um trotzdem einen Verlust von Privatsphäre zu erreichen: Ein Mensch kann ein Foto auch einer Person zuordnen, wenn dieses Foto Rauschen und Fehler aufweist.

Festgehalten werden muss auch, dass viele der Risiken im Kontext Privatheit und Maschinelles Lernen bisher erst wissenschaftliche Fragestellungen sind, deren theoretische Umsetzbarkeit durch Experimente belegt wurde. Beispiele aus der Praxis, die bekannt geworden sind, fehlen hier noch. Allerdings sollte Sicherheitsplanung immer auch zukünftige Risiken einbeziehen. Das gilt besonders dann, wenn Handlungen, die zu einem Verlust von Privatheit führen, nicht umkehrbar sind. Bei einem Leakage Problem beispielsweise sind die Daten im Nachhinein nicht löschbar. Soll also beispielsweise ein trainiertes Netz veröffentlicht werden, erfordert ein verantwortungsbewusster Umgang mit personenbezogenen Daten bei deren Nutzung bereits ein Abwägen der Risiken der hier erörterten Angriffe. Perspektivisch muss immer davon ausgegangen werden, dass Angriffe, die theoretisch denkbar sind, irgendwann auch in der Praxis umgesetzt werden.

7.4 Literaturverzeichnis

[1] Fredrikson, M., Jha, S. and Ristenpart, T., 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, S. 1322-1333.

[2] Shokri, R., Stronati, M., Song, C. and Shmatikov, V., 2017. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy, S. 3-18.

[3] Döbel, I., Leis, M., Vogelsang, M. and Petzka, H., 2018. Maschinelles Lernen. Eine Analyse zu Kompetenzen. Forschung und Anwendung. Fraunhofer-Gesellschaft, München.

[4] Shwartz-Ziv, R. and Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.

[5] Al-Rubaie, M. and Chang, J.M., 2019. Privacy-preserving machine learning: Threats and solutions. IEEE Security & Privacy, 17(2), S. 49-58.

[6] Faception – Facial Personality Analysis, **↗https://www.faception.com/**.

[7] Zhang A, Lipton ZC, Li M, Smola AJ, 2019. Dive into Deep Learning. **↗https://d2l.ai/index.html**.

[8] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A., 2017. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia conference on computer and communications security, S. 506-519.

[9] Wang, B. and Gong, N.Z., 2018. Stealing hyperparameters in machine learning. In 2018 IEEE Symposium on Security and Privacy, S... 36-52.

[10] Duddu, V., Samanta, D., Rao, D.V. and Balas, V.E., 2018. Stealing neural networks via timing side channels. arXiv preprint arXiv:1812.11720.

[11] Batina, L., Bhasin, S., Jap, D. and Picek, S., 2018. CSI neural network: Using side-channels to recover your artificial neural network information. arXiv preprint arXiv:1810.09076.

[12] Hua, W., Zhang, Z. and Suh, G.E., 2018. Reverse engineering convolutional neural networks through side-channel information leaks. In 2018 55th ACM/ESDA/IEEE Design Automation Conference, S. 1-6.

[13] Pal, S., Gupta, Y., Shukla, A., Kanade, A., Shevade, S. and Ganapathy, V., 2019. A framework for the extraction of deep neural networks by leveraging public data. arXiv preprint arXiv:1905.09165.

[14] Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A. and Papernot, N., 2019. High-fidelity extraction of neural network models. arXiv preprint arXiv:1909.01838.

[15] Correia-Silva, J.R., Berriel, R.F., Badue, C., de Souza, A.F. and Oliveira-Santos, T., 2018. Copycat CNN: Stealing knowledge by persuading confession with random non-labeled data. In 2018 International Joint Conference on Neural Networks, S. 1-8.

[16] Mosafi, I., David, E.O. and Netanyahu, N.S., 2019. Stealing knowledge from protected deep neural networks using composite unlabeled data. In 2019 International Joint Conference on Neural Networks, S. 1-8.

[17] Oh, S.J., Schiele, B. and Fritz, M., 2019. Towards reverse-engineering black-box neural networks. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 121-144. Springer, Cham.

[18] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[19] Dwork, C., 2009. The differential privacy frontier. In Theory of Cryptography Conference, S. 496-502. Springer, Berlin, Heidelberg.

[20] Dwork, C., McSherry, F., Nissim, K. and Smith, A., 2006. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference, S. 265-284. Springer, Berlin, Heidelberg.

[21] Page, H., Cabot, C. and Nissim, K., 2018. Differential privacy: an introduction for statistical agencies. NSQR. Government Statistical Service.

[22] Definition Kryptographie, 2017. **↗https://www.securityinsider.de/was-ist-kryptographie-a-642288/**.

[23] Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M. and Wernsing, J., 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International Conference on Machine Learning, S. 201-210.

[24] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S. and Leung, V.C., 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. IEEE access, 6, S. 12103-12117.

[25] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K., 2017. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, S. 1175-1191.

[26] Hitaj, B., Ateniese, G. and Perez-Cruz, F., 2017. Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, S. 603-618.

[27] McMahan, H.B., Moore, E., Ramage, D. and Hampson, S., 2016. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629.

[28] Melis, L., Song, C., De Cristofaro, E. and Shmatikov, V., 2019. Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE Symposium on Security and Privacy, S. 691-706.

[29] Shokri, R. and Shmatikov, V., 2015. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security S. 1310-1321.

[30] Li, H., Xiong, L. and Jiang, X., 2014. Differentially private synthesization of multi-dimensional data using copula functions. In Advances in database technology: proceedings. International conference on extending database technology, Vol. 2014, S. 475. NIH Public Access.

[31] Triastcyn, A. and Faltings, B., 2018. Generating artificial data for private deep learning. arXiv preprint arXiv:1803.03148.

[32] Bowen, C.M. and Liu, F., 2016. Comparative study of differentially private data synthesis methods. arXiv preprint arXiv:1602.01063.

8 Anonymisierung und Pseudonymisierung medizinischer Textdaten mittels Natural Language Processing

8 Anonymisierung und Pseudonymisierung medizinischer Textdaten mittels Natural Language Processing

Benedikt Kämpgen, Dominic Swarat

Medizinische Datenanalysen, Natural Language Processing, unstrukturierte Daten, Maskierung, Anonymisierung

Im eHealth-Bereich drehen sich Anwendungsfälle häufig um das Bereitstellen von Gesundheitsdiensten auf Basis der Verwendung von Patientendaten. Hierbei werden Datenbanken oder Dateisysteme genutzt, um entsprechende Patientendaten abzufragen.

Zum Beispiel sind in der Radiologie die Alt-Befunde die Grundlage für eine retrospektive Analyse [17,18]. Der Befund, unmittelbar während er geschrieben oder diktiert wird, ist die Grundlage für prospektive Entscheidungsunterstützung [19].

Für viele Anwendungsfälle muss daher jeweils ein durchdachtes **Datenschutzkonzept** vorhanden sein [1]. Datenschutz bedeutet, den Einzelnen davor zu schützen, dass er durch den Umgang mit seinen personenbezogenen Daten in seinem Recht auf informationelle Selbstbestimmung beeinträchtigt wird. Es gibt eine Vielzahl an Regularien, Normen, Standards für den Datenschutz.

Beim Umgang mit personenbezogenen Daten gilt, dass »die Pseudonymisierung und Verschlüsselung personenbezogener Daten« eine geeignete technische und organisatorische Maßnahme darstellt, »um ein dem Risiko angemessenes Schutzniveau zu gewährleisten« [2] zur erforderlichen Pseudonymisierung und Verschlüsselung personenbezogener Daten).

Auch bei einer »Datenverarbeitung zu wissenschaftlichen oder historischen Forschungszwecken und zu statistischen Zwecken« [3] sind personenbezogene Daten häufig zu anonymisieren.

Eine Definition von **Anonymisierung** laut ISO 29100:2011 ist: »Anonymization is the process by which personally identifiable information (PII) is irreversibly altered in such a way that a PII principal can no longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party.«

Man kann zwischen Anonymisierung von **strukturierten Daten** und **unstrukturierten Daten** unterscheiden.

Für strukturierte Daten nutzen wir die Definition aus ↗**Kapitel 2**: »In diesem Leitfaden sprechen wir von Datensätzen, welche aus einzelnen Datenpunkten bestehen. Jeder Datenpunkt (oder Zeile) des Datensatzes enthält Attribute, die konkrete Werte besitzen«. Man kann hierbei also von tabellarischen (oder sogar relationalen) Daten mit kategorischen Werten ausgehen.

Bei der Anonymisierung strukturierter Daten können konkrete Werte für jedes Attribut angenommen werden. Das heißt, jeder Wert lässt sich einer Kategorie zuordnen, ist kategorisch.

Mit unstrukturierten Daten werden typischerweise Daten bezeichnet, die kein festes Schema aus Tabellenspalten und kategorischen Werten besitzen, z. B. Zeitreihen, Medieninhalte wie Bild, Ton und Video (siehe ↗Kapitel 7) sowie Freitext. Unter Freitext verstehen wir mit Tastatur oder Diktiergerät (und Spracherkennung) geschriebene natürliche Sprache.

Unter unstrukturierten Daten versteht man primär Freitextdokumente, Bilder sowie Ton-/Videoaufnahmen. Auch Gerätesignaldaten werden manchmal darunter verstanden, da es sich um Rohdaten handelt, die durch ihren spezifischen Verlauf auf einzelne Personen schließen lassen können.

Bilder, Ton-/Videoaufnahmen und Gerätesignaldaten sind sehr speziell und werden in diesem Dokument nicht behandelt. Der Großteil der auswertbaren Informationen stecken heutzutage in Krankenhäusern in **Freitextdokumenten**, z. B. Arztbriefe, radiologische Befundberichte, OP-Berichte.

Häufig liegen gerade medizinische Daten nicht komplett strukturiert in Datenbanken vor, sodass eine Anonymisierung/Pseudonymisierung mittels bewährter Methoden nicht direkt anwendbar ist.

Das heißt, ein Großteil medizinischer Daten, die für Maschinelles Lernen nützlich sind, sind semi-strukturiert, d. h. einzelne Attribute können beliebig langen natürlichsprachlichen Freitext enthalten.

Hier können Methoden zur Anonymisierung von strukturierten Daten nicht unmittelbar angewendet werden, zumindest nicht, ohne einen hohen Informationsverlust.

Dennoch kann man auch sicherstellen, dass natürlichsprachlicher Freitext anonym ist.

Wir unterscheiden primär drei Möglichkeiten, die in den folgenden Kapiteln jeweils beschrieben werden.

8.1 Anonymisieren im Voraus

Hier wird im Voraus sichergestellt, dass Freitext keine identifizierenden Begriffe enthalten. Dies ist durch technisch-organisatorische Maßnahmen möglich, z. B. einen eindeutigen Hinweis an dateneingebende Personen. Dies ist zum Beispiel häufig bei radiologischen Befundberichten der Fall. Oder bei den Feldern für »Diagnosen« in Arztbriefen.

Personenbezogene Daten sind alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen, beispielsweise Name, Vorname, Geburtsdatum, Lohn/

Gehalt, Familienstand. Besonders sensible Daten sind Angaben über die rassische und ethnische Herkunft, politische Meinung, religiöse oder philosophische Überzeugung, Gewerkschaftszugehörigkeit, Gesundheit oder das Sexualleben.

8.2 Anonymisierung durch Maskierung

Hier wird eine nachträgliche Maskierung von identifizierenden Merkmalen vorgenommen. Durch Analyseverfahren können entsprechende Merkmale extrahiert und entfernt bzw. durch Platzhalter ersetzt werden.

Dies kann manuell geschehen: Bei kleineren Anonymisierungstätigkeiten auf allgemeinem Text hat sich ein guter PDF-Editor mit Schwärzungsfunktion als praktisch erwiesen (z. B. Foxit).

Allerdings ist der Aufwand sehr hoch, daher können auch automatische Verfahren bei der Extraktion von identifizierenden Merkmalen genutzt werden.

Die Herausforderung bei der Anonymisierung von Freitextdaten besteht darin, dass sie identifizierende Merkmale des Patienten enthalten, aber nicht durch einfache Operationen auf Merkmalen anonymisiert werden können. Stattdessen müssen die identifizierenden Merkmale (ob allein oder in Kombination) zunächst gefunden werden, was aufgrund verschiedener Schreibweisen nicht einfach ist.

So ist es zwar grundsätzlich möglich anhand von Metadaten wie Patientenname, Anschrift und Geburtsdatum diese anschließend in den Freitexten zu identifizieren und zu entfernen. Mittels individuell geführter Listen können bestimmte Textpassagen explizit entfernt (z. B. krankenhausspezifische Besonderheiten wie Arzt- und Stationsnamen) oder erhalten (z. B. Produktbezeichnungen oder Eigennamen wie Parkinson) werden. Auch gibt es Verfahren um Informationen wie Eigennamen (auch die von Angehörigen oder von niedergelassenen Ärzten), Orte, Email-Adressen und Datumsangaben automatisch zu erkennen.

Allerdings gibt es keine existierende Software, die zu 100% zuverlässig Freitextdaten anonymisiert. Es ist immer noch ein Prüfen durch Menschen notwendig, mindestens stichprobenartig, eigentlich jedes einzelnen Dokuments. Aus diesem Grund bringen manche solcher Tools insgesamt mehr Overhead als Nutzen, insbesondere wenn es darum geht, eingescannte Dokumente zunächst mittels OCR elektronisch aufzubereiten.

8.3 Anonymisierung durch Natural Language Processing

Hier werden durch Methoden des Natural Language Processing Freitextdaten strukturiert und anschließend auf diesen strukturierten Daten herkömmliche Methoden zu Anonymisierung von Freitextdaten angewendet.

Folgende Abbildung zeigt schematisch die Umsetzung bei der Pseudonymisierung bzw. Anonymisierung.

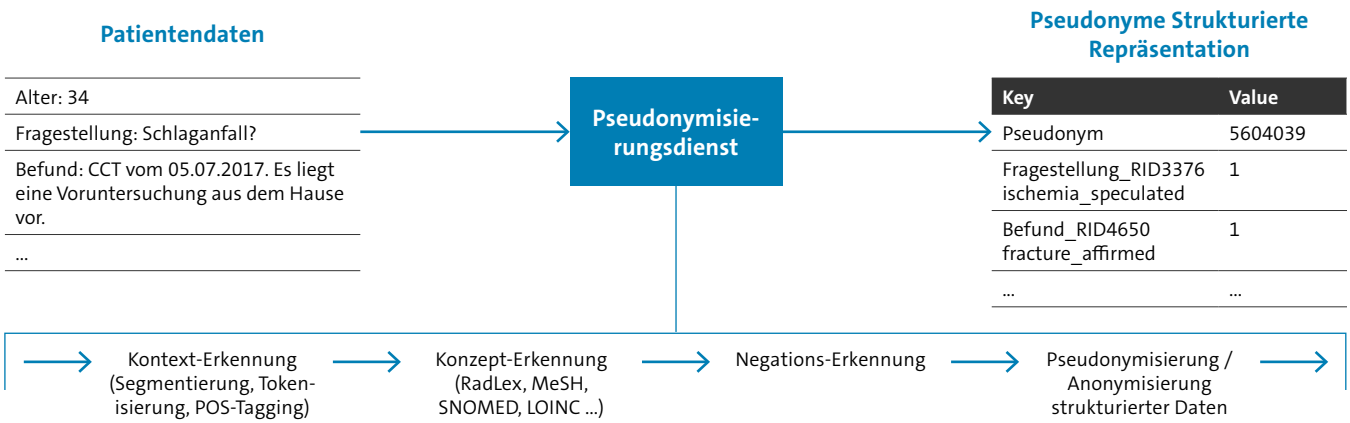

Abbildung 12: Umsetzungsvorschlag »Pseudonymisierungsdienst«

Hierbei werden Patientendaten mittels eines »Pseudonymisierungsdienst« pseudonymisiert. Dabei werden die folgenden Schritte unternommen:

Strukturierung

Freitextfelder in Patientendaten werden mittels Natural Language Processing (NLP) zunächst in eine strukturierte Form gebracht:

- Kontext-Erkennung: Der Kontext einer Information, z. B. das ursprüngliche Dokument, wird erkannt (z. B. »Entlassung_Neuro«, »MRT«)
- Konzept-Erkennung: Konzepte basierend auf öffentlich verfügbaren Terminologien wie RadLex, MeSH, ICD, OPS etc. werden extrahiert
- Negations-Erkennung: Konzepte werden weiter beschrieben, z. B. in »bestätigt«, »ausgeschlossen«, »vermutet«
- Beliebige weitere NLP-Aufgaben können durchgeführt werden, z. B. temporale Erkennung, quantitative Erkennungen

Für solches NLP nutzen wir die Empolis Healthcare Analytics Services [16], es gibt allerdings auch andere Open-Source oder kommerzielle Werkzeuge, die hierfür genutzt werden können.

Die einzelnen Merkmale sind per Definition nicht personenbezogen, weil es sich um Standard-Konzepte öffentlich zugänglicher Terminologien handelt.

Lediglich Kombinationen von einzelnen Merkmalen könnten eine Re-identifizierung ermöglichen; dies kann verhindert werden, indem anschließend die Menge aller generierten strukturierten Repräsentationen mittels Ansätze für strukturierte Daten anonymisiert und mittels k-Anonymitätskriterium überprüft werden.

Anonymisierung

Ein häufiger Ansatz besteht in der sog. **De-Identifizierung**, bei der alle identifizierenden Merkmale herausgesucht und ersetzt werden.

In den USA laut HIPAA Safe Harbor [6],[7] ist darunter die Entfernung bzw. Ersetzung folgender Informationen gemeint: Namen, Geografische Unterteilungen, inklusive Adressen, Datumswerte speziell für einzelne Personen, z. B. Geburtsdatum, Telefonnummern, Faxnummern, E-Mail-Adressen, Sozialversicherungsnummern, Nummer der Krankenakte, Krankenversicherungsnummern, Kontonummern, Zertifikats- oder Lizenznummern, Fahrzeugkennzeichen, Geräteidentifikationsnummern, URLs, IPs, biometrische Identifikatoren, Vollbildfotos und vergleichbare Bilder, andere eindeutige Identifikationsnummern.

Letztendlich handelt es sich hierbei um die Anonymisierungstechnik »Generalisierung« [8], bei der die verbleibenden Informationen es nicht erlauben, einzelne Personen zu identifizieren (bspw. das Geburtsdatum in das allgemeinere Geburtsjahr umgeschrieben wird).

Es gibt eine Vielzahl weiterer Techniken, wie z. B. das Hinzufügen von Rauschen [9] oder das modellbasierte synthetische Generieren von Daten [15].

Für die Anonymisierung strukturierter Daten gibt es bewährte Werkzeuge, z. B. das ARX Data Anonymization Tool [12].

Pseudonymisierung

Zu unterscheiden ist Anonymisierung und **Pseudonymisierung**. Während es bei der Anonymisierung für den Verantwortlichen vernünftigerweise nicht möglich ist, Rückschlüsse auf Betroffene zu ziehen, ist dies bei der Pseudonymisierung weiterhin möglich. Bei der Pseudonymisierung erhält jeder Fall ein zusätzliches neu erfundenes identifizierendes Merkmal, ein Pseudonym, das es demjenigen, der die Zuordnung von Pseudonym zu ursprünglichen identifizierenden Merkmalen kennt, ermöglicht, den Fall wieder zuzuordnen.

Als Beispiel, laut dem Forschungsprojekt Theseus: In der Begrifflichkeit wurde in Anlehnung an TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. [4] für das Medico-Projekt eine Pseudonymisierung durchgeführt [5]. Dabei wurden die patientenidentifizierenden Daten von den zu exportierenden (Bild-)Daten entfernt und getrennt geführt, um

eine Rückverfolgung oder De-Pseudonymisierung durchführen zu können. Im Projekt verblieben diese patienten-identifizierenden Daten im medizinischen Netz der Universitätsklinik Erlangen. Es fand dabei für die genannten DICOM-Felder eine Namensänderung durch ein Pseudonym statt.

Daher wird anschließend den anonymisierten Informationen ein aus identifizierenden Merkmalen generiertes Pseudonym hinzugefügt.

Auch für die Erstellung von Pseudonymen gibt es Werkzeuge, z. B. PID-Generator[13] bzw. das Nachfolgeprodukt Mainzelliste[14].

8.4 Auswahl, Voraussetzungen der Anonymisierungsmethode

Bei der Anonymisierung von Freitextdatenfeldern hängt es u.a. von der Art der zu anonymisierenden Daten, dem geplanten Verwendungszweck der Daten sowie den technischen und organisatorischen Rahmenbedingungen der Datennutzung ab, welches Verfahren anwendbar ist. So kann es z. B. sein, dass der Autor nicht identifizierbar sein soll und daher auch der Schreibstil anonymisiert werden soll.

Die **Bewertung** der entsprechenden Anonymisierungstechnik geschieht nach zwei Gesichtspunkten: Möglichkeiten einer Re-Identifizierung und Anwendbarkeit für den Anwendungsfall, diese werden im Folgenden kurz beschrieben.

Möglichkeiten einer Re-Identifizierung: Auch nach einer De-Identifizierung kann eine Re-Identifizierung, also die tatsächliche Identifizierung einer einzelnen Person, z. B. durch die Kombination mit anderen Datenquellen, nie zu 100% ausgeschlossen werden; so muss im Datenschutzkonzept und durch die Technisch-Organisatorischen Maßnahmen sichergestellt werden, dass eine Re-Identifizierung nur zu einer geringen Wahrscheinlichkeit und nur durch höchsten Aufwand erreicht werden kann.

Dies kann mit weiteren Schutzmaßnahmen kombiniert werden: Zum Beispiel sollten Datenempfänger sich explizit verpflichten, keine Re-Identifizierung vorzunehmen. Außerdem kann durch eine Risikoanalyse aufgezeigt werden, dass auch im Fall einer Re-Identifizierung der Schaden für die betroffene Person höchstwahrscheinlich gering ist.

Eine Möglichkeit, um die Effizienz einer Anonymisierungstechnik zu bewerten ist K-Anonymität [10]. Dies bedeutet, dass es immer mindestens k Fälle gibt, die die gleichen Merkmale aufweisen.

Anwendbarkeit für den Anwendungsfall: Allerdings sollten Anonymisierungstechniken nicht nur danach bemessen werden, wie hoch die Genauigkeit beim De-Identifizieren ist, sondern auch, wie viel nützliche (aber anonyme) Informationen für die Datenanalyse im Anwendungsfall erhalten bleiben [11].

Voraussetzung für diesen Ansatz ist die grobe Definition der Anwendungsfälle für die anonymisierten Daten. Denn bei der Strukturierung mittels Standard-Konzepten und Natural Language Processing geht immer Information verloren; eine 1:1-Übersetzung in strukturierte Form ist nicht möglich. Daher muss im Vorfeld festgelegt werden, für welche Art von Use Cases und Analytics-Aufgaben die anonymisierten Daten hergenommen werden sollen.

So handelt es sich – je nach Anwendungsfall – bei allen drei Methoden um pragmatische Ansätze der effektiven Anonymisierung bzw. Pseudonymisierung von Patientendaten zum Maschinellen Lernen.

8.5 Literaturverzeichnis

[1] ↗**https://www.toolpool-gesundheitsforschung.de/produkte/checkliste-zur-erstellung-eines-datenschutzkonzeptes**
[2] ↗**https://dsgvo-gesetz.de/art-32-dsgvo/**, siehe DSGVO Sicherheit der Verarbeitung, hier besonders Satz 1 a
[3] ↗**https://dsgvo-gesetz.de/bdsg-neu/27-bdsg-neu/**, siehe DSGVO Datenverarbeitung zu wissenschaftlichen oder historischen Forschungszwecken und zu statistischen Zwecken hier besonders Satz 3, zur Anonymisierungspflicht und Satz 1 zur Erfordernis der personenbezogenen Daten
[4] ↗**http://www.tmf-ev.de/Home.aspx**
[5] ↗**http://www.egms.de/en/meetings/gmds2005/05gmds388.shtml**
[6] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., Samore, M. H., Dorr, D., … Uzuner, O. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology, 10(1), 70. ↗**https://doi.org/10.1186/1471-2288-10-70**
[7] Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., … Clifford, G. D. (2008). Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making, 8(1), 32. ↗**https://doi.org/10.1186/1472-6947-8-32**
[8] Directive 95/46/EC of the European Parliament. (2014). Opinion 05/2014 on Anonymisation Techniques, (April), 37. Retrieved from ↗**http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf**
[9] Directive 95/46/EC of the European Parliament. (2014). Opinion 05/2014 on Anonymisation Techniques, (April), 37. Retrieved from ↗**http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf**
[10] Directive 95/46/EC of the European Parliament. (2014). Opinion 05/2014 on Anonymisation Techniques, (April), 37. Retrieved from ↗**http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf**
[11] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., Samore, M. H., Dorr, D., … Uzuner, O. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology, 10(1), 70. ↗**https://doi.org/10.1186/1471-2288-10-70**
[12] ↗**https://arx.deidentifier.org/**

[13] ↗https://www.toolpool-gesundheitsforschung.de/produkte/pid-generator
[14] ↗https://www.toolpool-gesundheitsforschung.de/produkte/mainzelliste
[15] ↗https://www.datenschutzbeauftragter-info.de/synthetische-daten-die-rettung-aus-der-anonymisierungskrise/
[16] ↗https://www.empolis.com/empolis-healthcare-analytics-services/
[17] Jungmann, F. et al. Towards data-driven medical imaging using natural language processing in patients with suspected urolithiasis. International Journal of Medical Informatics (2020)
[18] Jungmann, F., Kuhn, S. & Kämpgen, B. Grundlagen und Einsatzmöglichkeiten von Natural Language Processing (NLP) in der Radiologie. Radiologe 58 (2018)
[19] Maros, M., Kämpgen, B., Förster, A., Groden, C., Sommer, W. & Klüter, A. Structured reporting supports junior readers and improves PI-RADS conformity of multi-parametric MRI reports of the prostate-based on cross-lingual RADLEX annotations. ECR (2018)

9 Semantische Anonymisierung sensibler Daten mit inferenzbasierter KI und aktiven Ontologien

9 Semantische Anonymisierung sensibler Daten mit inferenzbasierter KI und aktiven Ontologien

Bernd Geiger, Hermann Rapp, Narayanan Sampath

Semantische KI, Semantic Anonymisation, Aktive Ontologien, OntoBroker

In der Finanzindustrie sind vielfältige interne und externe datenschutzrechtliche Vorgaben und regulatorische Rahmenbedingungen zu beachten. Dieser Beitrag stellt Semantische Anonymisierung als eine neue Methode vor, die es ermöglicht, sensible Daten mit einem semantischen KI-basierten System von semafora systems mit aktiven Ontologien und Inferencing so zu verändern, dass sie datenschutzgerecht analysiert werden können. Bisherige Verfahren [1, 2, 3] nutzen entweder eine Pseudonymisierung durch kryptographische Verfahren und Tokens oder eine Anonymisierung durch eine Verzerrung der Daten bzw. durch Veränderung oder Entfernen von Details aus einem Datensatz.

Dieser neue KI-basierte Ansatz erhält im Gegensatz zu bisherigen Methoden weitgehend die Aussagekraft der Rohdaten und erlaubt eine datenschutzkonforme Analyse personenbezogener Daten, bei der gewährleistet ist, dass die ursprünglichen Personen nicht mehr identifiziert werden können und auch eine personenbezogene Rückverfolgbarkeit über Quasi-Identifiers verhindert wird (Zwei-Wege-Sicherung). Große Datenmengen mit sensiblen Daten (personenbezogen bzw. unternehmensstrategisch) können durch diese Methode datenschutzgerecht, aber zugleich unter Erhaltung des Analysepotenzials, genutzt werden.

In dem Beitrag werden zwei Fallbeispiele aus der Finanzindustrie dargestellt, wobei die KI-Systemlösung auch in anderen Industrien anwendbar ist. Die vorgestellten Anwendungsfälle sind, erstens, Analysedaten, die zur Produktentwicklung bzw. Optimierung von Marketing-Kampagnen und einer verbesserten Kundenansprache genutzt werden. Zweitens lassen sich Testdaten aus Echtdaten generieren, mit denen realistische Tests von Schnittstellen bzw. IT-Systemen möglich sind. Bestandteil von Semantischer Anonymisierung als Methode ist eine Dokumentation für Auditzwecke.

9.1 Aktive Ontologien – die nächste Generation

Im Folgenden wird kurz die nächste Generation der bislang bekannten (konventionellen) Ontologien, die sogenannten aktiven Ontologien, beschrieben. Mit **konventionellen Ontologien** lassen sich Daten und deren Zusammenhänge abbilden, z. B. das Wissen, das in einer Taxonomie-orientierten Struktur bestehend aus Klassen, Unterklassen, Attributen und Relationen gespeichert ist. Konventionelle Ontologien sind gerichtete Graphen und zur Datenspeicherung flexibler als relationale Datenbanken.

Aktive Ontologien sind Ausführungsumgebungen (Semantic Runtime Environments) und verwenden logische Programmierung mit Funktionen für logisch-funktionale Zusammenhänge. Sie können aktiv Aktionen ausführen, z. B. Datentransformationen und ereignisgesteuerte Smart Contracts. Der Vorteil von aktiven Ontologien im Vergleich mit konventionellen Ontologien ist es, dass nicht nur abstrahiertes Wissen und die dazugehörigen Instanzen (die eigentlichen Daten, z. B. die Menge aller Nachnamen) beinhaltet sind, sondern dass sich die Inhalte der Ontologie (also alle Klassen, Instanzen Eigenschaften und Relationen) dynamisch mit Ontologie-Funktionen verändern lassen. So lassen sich adaptiv Daten aus der Ontologie zur Anonymisierung verändern unter Beibehaltung einer maximal möglichen Ähnlichkeit der anonymisierten Daten mit den Echtdaten.

9.2 Semantische Technologie und industrielle Einsatzmöglichkeiten

Als Semantic Runtime Environment (Inference Engine) wird OntoBroker von semafora systems eingesetzt, der seine Stärken insbesondere in der Prozessierung der funktionalen Aspekte der Ontologien hat. Das skalierbare System wurde in mehr als zwei Jahrzehnten in Zusammenarbeit mit Industriekunden zu industrieller Performanz optimiert. Die Architektur und der Datendurchsatz sind so ausgelegt, dass große Datenmengen, die in der Finanzindustrie typischerweise vorkommen, performant verarbeitet werden. In Verbindung mit aktiven Ontologien und Higher Order Logic (HOL) [4] lässt sich Semantische Anonymisierung im industriellen Maßstab realisieren.

Besondere Werkzeuge und Bibliotheken (taksai Data Technologies) für die Vor- und Aufbereitung der Daten zum Import in OntoBroker und der Analyse von Daten werden bereitgestellt.

9.3 Semantische Anonymisierung

Semantische Anonymisierung ist eine neuartige Methode, mit der personenbezogene Daten so anonymisiert werden können, dass gem. DS-GVO (Erwägungsgrund 26) für die Verarbeitung solcher anonymisierten Daten, bei denen betroffene Personen nicht oder nicht mehr identifiziert werden können, die DS-GVO ausdrücklich nicht gilt.

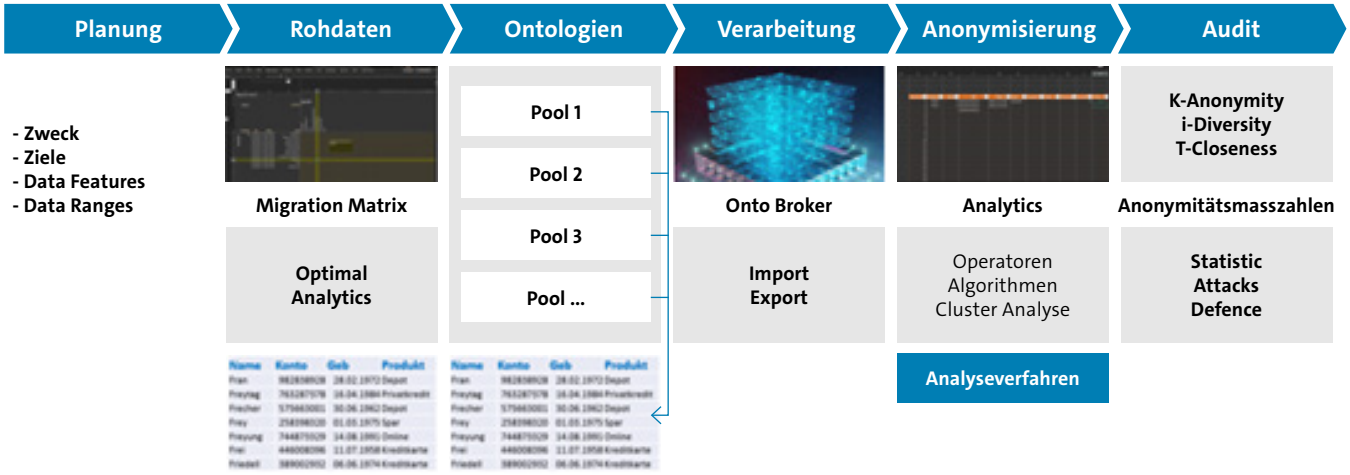

Abbildung 13: Abfolge der Schritte bei Semantischer Anonymisierung

Im Weiteren sind die einzelnen Schritte im Detail beschrieben.

Schritt 1: Planung, Vorbereitung und Datendesign

In einem ersten Schritt beginnt unter Einbeziehen von Domainexperten das Verfahren mit einer Planung und Definition des Nutzungszwecks und der Analyseziele sowie der Parameter, die für die geplanten Analysen relevant sind. Teil der Planung ist auch eine Einstufung bezüglich IT- und Datensicherheit und Festlegung der notwendigen Maßnahmen. Dabei kann auch eine Klassifizierung einzelner Attribute bzw. Datenelemente vorgenommen werden, z. B. aufgrund unternehmensspezifischer oder regulatorischer Datenschutzvorgaben.

Schritt 2: Statistische Analyse der Rohdaten

Abhängig von Nutzungszweck und Analysezielen sowie der Art der Daten können gruppenbezogene Verhältnisse und Muster in den Rohdaten berechnet und im weiteren Prozess erhalten werden. Dies wird durch verschiedene deskriptive und analytische statistische Methoden erreicht. Im Gegensatz zu aktuellen Techniken wie Differential Privacy [1, 4, 5], bei der Analyseergebnisse mit fast gleicher Wahrscheinlichkeit erzielt werden, lassen sich mit Semantischer Anonymisierung statistisch valide Aussagen nicht nur über eine gesamte Datenpopulation, sondern auch über bestimmbare Teilmengen treffen, jeweils abhängig von den Analysezwecken.

In Bezug auf die verwendeten Rohdaten ist ein Recht auf Löschung (Art. 17 DS-GVO) zu gewährleisten, das beinhaltet, dass alle Daten, die sich auf eine einzelne Person beziehen, auf deren Wunsch hin gelöscht werden können.

Schritt 3: Transformation der Daten unter Einsatz von aktiven Ontologien

Der dritte Schritt beinhaltet die Erstellung der Transformationsregeln und aktiven Ontologien, mit denen die Daten je nach Zielsetzung (Schritt 1) anonymisiert werden. Der Datenbestand wird dabei in einzelne Datenpools aufgeteilt, die je nach logisch-funktionalem Zusammenhang unterschiedlich behandelt werden. Die Aufteilung in Gruppen (Datenpools) unter Verwendung von aktiven Ontologien erlaubt eine getrennte Verarbeitung der Variablen (Semantische Mikroaggregation), was die Bandbreite der Abweichungen zwischen Rohdaten und anonymisierten Daten relativ gering hält.

Handelt es sich beispielsweise um Daten für eine Analyse der räumlichen Verteilung bestimmter Personengruppen, so kann durch Geospatial Semantics festgelegt werden, ob geographische Angaben wie Adressdaten (Straße, Ort, Land) innerhalb von bestimmten räumlichen Grenzen verändert werden. Im Gegensatz zu Differential Privacy Methoden [5] werden dabei die Daten nicht zufällig geändert, sondern unter Erhaltung der in Schritt 2 generierten Verteilungseigenschaften zwischen den Datensubjekten. Dies geschieht innerhalb zu spezifierender Gruppen (Data Pools), die je nach Analysezielen differieren können. Ein eventueller individueller Personenbezug wird durch die Bildung ausreichend großer Datenpools vermieden.

Analog gilt dies für zeitliche oder produktbezogene Parameter und sozio-ökonomische Dimensionen wie Berufs- oder Einkommensgruppen. Beispielsweise kann der Analysezweck darin bestehen, den zeitlichen Aspekt zwischen der Eröffnung eines Online-Kontos und der Nutzung bestimmter Produkte in Abhängigkeit von bestimmten Einkommensgruppen zu untersuchen. Weitere Parameter für die Analyse können dabei die Sprache sein, die Kunden für die Kommunikation mit dem Finanzinstitut ausgewählt haben, oder die Kanäle, über die Kunden kommunizieren möchten.

Unternehmensspezifische Daten wie z. B. Kontonummern werden unter Einhaltung der Datenformate mit entsprechenden Algorithmen so verändert, dass auch in den anonymisierten fiktiven Kontonummern gültige Prüfziffern enthalten sind.

Schritt 4: Datenverarbeitung mit der Inference Engine und Generierung der anonymisierten Daten

Die Transformationsfunktionen und Regeln werden über MS Excel Templates erstellt, die dann in die Inference Engine importiert werden und automatisch in die Transformationsontologie gewandelt werden. Danach kann der Datenbestand (Echtdaten) in die Inference Engine importiert werden. Dort werden die Daten transformiert und als anonymisierter Datenbestand bereitgestellt zum Export in die Analyseumgebung.

Schritt 5: Datenanalyse

Die anonymisierten Daten können – abhängig von den Analysezielen – durch Auswahl aus einer Vielzahl von Modellen und Algorithmen sowie mit verschiedenen statistischen Verfahren

analysiert werden. Operatoren und Algorithmen können dafür aus einer Bibliothek abgerufen werden. Für besondere Analysezwecke können weitere Operatoren und Algorithmen erstellt bzw. importiert werden.

Schritt 6: Auditfähige Dokumentation

In der Finanzindustrie sind hohe regulatorische Anforderungen zu erfüllen. Dafür finden interne und externe Audits statt, die durch die auditfähige Dokumentation Semantischer Anonymisierung unterstützt werden.

Eine Rechtmäßigkeit der Verarbeitung gem. Art. 6 DS-GVO ist mit Bezug auf Training, Nutzung und Lebenszeit der verwendeten Ontologien und der Rohdaten sicherzustellen und zu dokumentieren.

Eine Identifizierung von individuellen Personen auch in Kombination verschiedener Parameter wie Beruf, Alter oder Wohnort (Personenbezug durch Quasi Identifiers) darf gem. DS-GVO nicht möglich sein. Dies wird dadurch erreicht, dass diese Ausprägungen innerhalb vorher zu bestimmender Gruppeninhalte semantisch transformiert und innerhalb der Datenpools zufällig gesetzt werden. Anonymitätskennzahlen (K-Anonymität, i-Diversität und T-Nähe) werden errechnet und dokumentiert.

9.4 Fallbeispiel 1: Analysedaten

Das erste Fallbeispiel zeigt die Anwendung der Semantischen Anonymisierung zur Erzeugung von Analysedaten aus Echtdaten, die aus Datawarehouse bzw. Business Intelligence Umgebungen des Finanzinstituts stammen.

Use Case 1: Analyse von Produktreichweite und Produktnutzung

Hier werden Transaktionsdaten zur Analyse von Produktreichweite und Produktnutzungsmustern von Finanzprodukten genutzt. Basis für die Analyse sind Ansätze zur Produktentwicklung bzw. Optimierung von Marketing-Kampagnen. Zweck der Analyse ist es, eine verbesserte Kundenansprache zu erreichen.

Use Cases 2: Analyse von Stammdaten

Analog zu Transaktionsdaten können auch Daten aus Stammdatensystemen zur Produktentwicklung bzw. Optimierung von Marketing-Kampagnen analysiert werden. Ziele dieser Analyse können eine verbesserte regionale Positionierung, Optimierung der Nutzung von Kommunikationskanälen und eine Erhöhung der Kundenzufriedenheit sein.

Entscheidend dafür ist, dass bereits in der Planung (Schritt 1) die Analysezwecke geklärt und notwendige Details mit Domainexperten und Vertretern des Analyseteams und der Fachseite (Business Units) vereinbart werden.

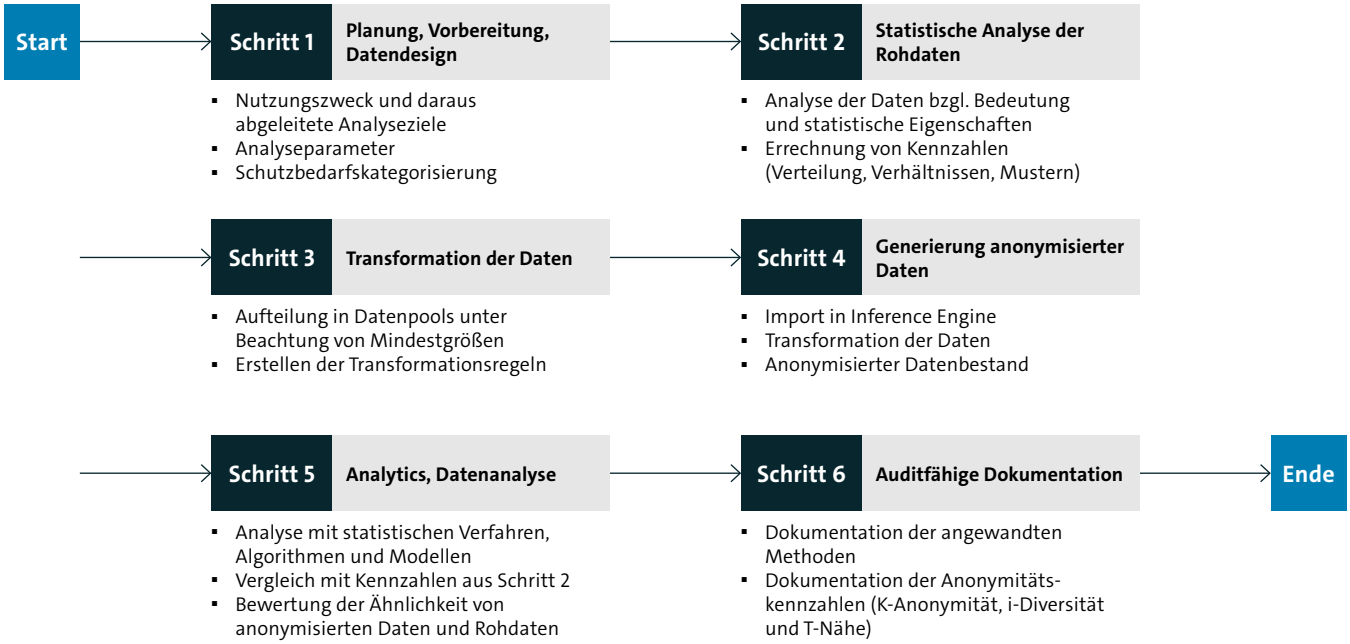

Abbildung 14: Ablauf bei Semantischer Anonymisierung für Analysedaten

Das Besondere ist dabei, dass durch den Prozess der Semantischen Anonymisierung ein anonymisierter Datenbestand bereitgestellt wird, der für zu spezifizierende Personengruppen (Schritt 1) bzw. Produktdaten einzelne Attribute und Parameter, die für die weitere Analyse entscheidend sind, gemäß ihren logisch-funktionalen Zusammenhängen und statistischen Verhältnissen bewahrt.

9.5 Fallbeispiel 2: Testdaten

Das zweite Fallbeispiel erklärt die Generierung von Testdaten aus Echtdaten. Aufgrund der datenschutzrechtlichen Regulatorik ist es nicht erlaubt, in Testsystemen Produktionsdaten zu verwenden oder dritten Parteien zur weiteren Verarbeitung ohne weiteres bereitzustellen. Deswegen müssen Testdaten generiert werden, um beispielsweise das Verhalten von Systemen zu testen.

Use Case: Realistische Tests von Schnittstellen bzw. IT-Systemen

Üblicherweise ist die Grundlage dafür eine Spezifikation der Schnittstelle und ein Testkonzept, in dem die Testobjekte definiert und die Testszenarien beschrieben sind. Dabei können je nach Testfall unterschiedliche Datenfelder und Daten spezifiziert werden. Dabei werden Werkzeuge bereitgestellt, um Daten je nach Spezifikation anpassen (Schritt 3) und verteilen (Schritt 5) zu können.

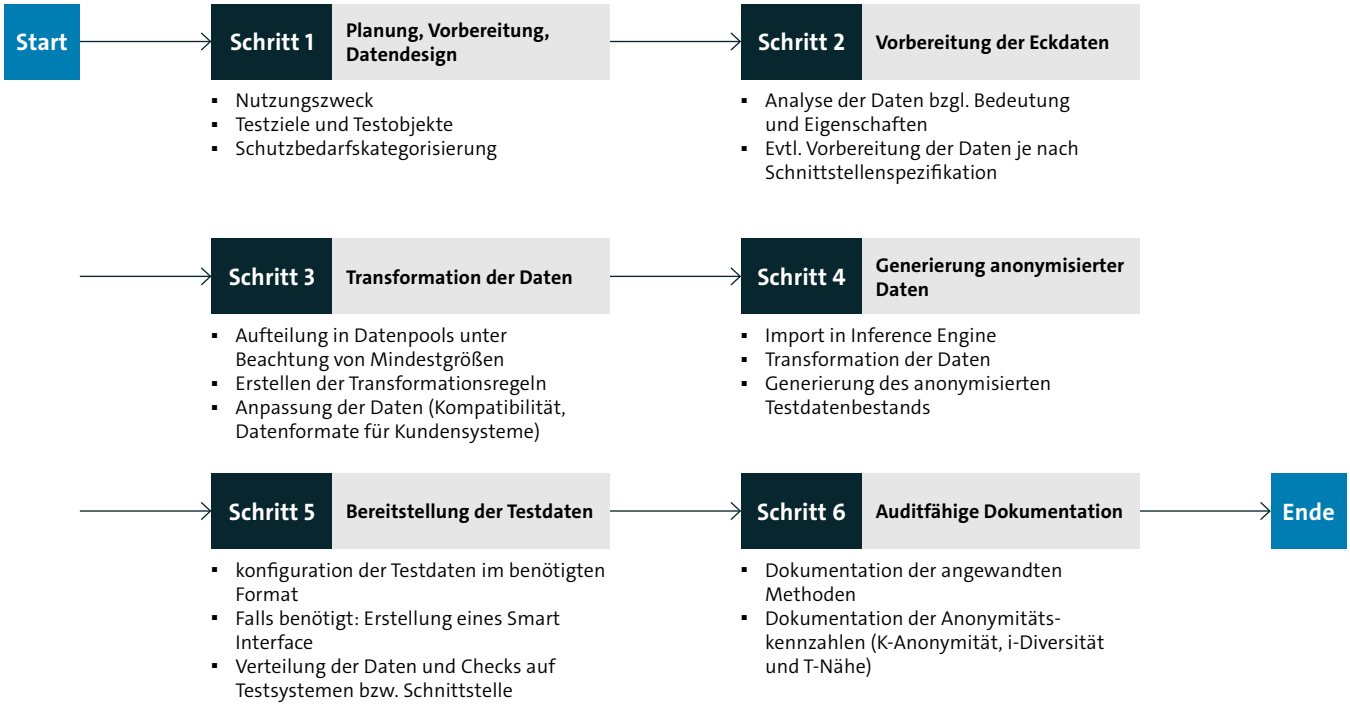

Abbildung 15: Ablauf bei Semantischer Anonymisierung für Testdaten

Entsprechend der oben beschriebenen Schrittfolge werden durch Semantische Anonymisierung aus Produktionsdaten anonymisierte Testdaten gewonnen, mit denen realistische Tests möglich werden.

9.6 Bewertung und Auditfähigkeit

Die Qualität einer Anonymisierung ist danach zu bewerten, inwieweit das Analysepotential eines Datenbestandes durch die Datenveränderung möglichst weitgehend erhalten bleibt. [6, 7]

Technisch geschieht dies bei der Semantischen Anonymisierung dadurch, dass Rohdaten bzw. Echtdaten analysiert werden (Schritt 2) in Bezug auf Bedeutung (Semantik) und statistische Verteilungseigenschaften und Verhältnisse sowie Muster zwischen verschiedenen Datenpunkten. Je nach Analysezielen werden Datenpools gebildet (Schritt 3), für die mit Hilfe von aktiven Ontologien Transformationsfunktionen erstellt werden. Die Generierung des anonymisierten Datenbestandes erfolgt durch Verarbeitung in der Inference Engine (Schritt 4). In Schritt 5 kann dann die eigentliche Analyse der anonymisierten Daten erfolgen, deren Güte durch Vergleich mit den in Schritt 2 errechneten Kennzahlen gemessen wird.

Die Dokumentation und auditfähige Beweisführung (Schritt 6) soll überprüfbar zeigen, dass Daten datenschutzgerecht für industrielle Analyse- bzw. Testzwecke verarbeitet und statistische Angriffe [8] verhindert werden. Semantische Anonymisierung als neue KI-Methode beinhaltet Tests, mit denen beweisbar ist, dass eine personenbezogene Rückverfolgbarkeit der Daten unmöglich ist (Zwei-Wege-Sicherung). Die Dokumentation dieser Tests dienen als Beweismittel für Auditzwecke. Als Maßzahlen für die Qualität der Anonymisierung und als statistische Sicherheitsbeweise werden K-Anonymität, i-Diversität und T-Nähe errechnet.

Zusammenfassend ist Semantische Anonymisierung eine innovative semantische KI-Methode zur Erreichung einer maximal möglichen Ähnlichkeit der anonymisierten Daten im Vergleich mit den Echtdaten.

9.7 Literaturverzeichnis

[1] Bitkom (2018) Machine Learning und die Transparenzanforderungen der DS-GVO. Leitfaden. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (bitkom), Berlin. Download über ↗**www.bitkom.org**

[2] Schwartmann, R. und Weiss, S. (Hrsg.) (2017) Whitepaper zur Pseudonymisierung. Leitlinien für die rechtssichere Nutzung von Pseudonymisierungslösungen unter Berücksichtigung der Datenschutz-Grundverordnung. Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2017.

[3] Schwartmann, R. und Weiss, S. (Hrsg.) (2019) Entwurf für einen Code of Conduct zum Einsatz DS-GVO konformer Pseudonymisierung. Arbeitspapier der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2019.

[4] Angele, J., Kifer, M. und Lausen, G. (2009). Ontologies in F-Logic. In: Staab, S. and Studer, R. (Eds.) Handbook on Ontologies, Second edition. Springer-Verlag, Berlin/ Heidelberg. Seiten 45-70.

[5] Ohm, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, Vol. 57, 1701-1777.

[6] Hundepool, A. et al. (2012) Statistical Disclosure Control. John Wiley & Sons, Ltd., Hoboken (NJ).

[7] Domingo-Ferrer, J. D. (2006) Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15: 355-369.

[8] Dziegielewska, O., Szafranski, B. (2016) A brief overview of basic inference attacks and protection controls for statistical databases. *Computer Science and Mathematical Modelling*, No. 4, 19-24.

Bitkom vertritt mehr als 2.700 Unternehmen der digitalen Wirtschaft, davon gut 2.000 Direktmitglieder. Sie erzielen allein mit IT- und Telekommunikationsleistungen jährlich Umsätze von 190 Milliarden Euro, darunter Exporte in Höhe von 50 Milliarden Euro. Die Bitkom-Mitglieder beschäftigen in Deutschland mehr als 2 Millionen Mitarbeiterinnen und Mitarbeiter. Zu den Mitgliedern zählen mehr als 1.000 Mittelständler, über 500 Startups und nahezu alle Global Player. Sie bieten Software, IT-Services, Telekommunikations- oder Internetdienste an, stellen Geräte und Bauteile her, sind im Bereich der digitalen Medien tätig oder in anderer Weise Teil der digitalen Wirtschaft. 80 Prozent der Unternehmen haben ihren Hauptsitz in Deutschland, jeweils 8 Prozent kommen aus Europa und den USA, 4 Prozent aus anderen Regionen. Bitkom fördert und treibt die digitale Transformation der deutschen Wirtschaft und setzt sich für eine breite gesellschaftliche Teilhabe an den digitalen Entwicklungen ein. Ziel ist es, Deutschland zu einem weltweit führenden Digitalstandort zu machen.

Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e.V.

Albrechtstraße 10
10117 Berlin
T 030 27576-0
F 030 27576-400
bitkom@bitkom.org
www.bitkom.org

bitkom