Herausforderungen für die Anonymisierung von Daten*

Christian Winter[1] Verena Battis[1] Oren Halvani[1]

Abstract: Unternehmen, Wissenschaftler und staatliche Stellen haben ein großes Interesse, neue Erkenntnisse aus Daten zu gewinnen. Dabei müssen Datenschutzregeln eingehalten werden. Anonymisierung ist auf den ersten Blick eine attraktive Lösung, um Datenschutz und Analyseinteressen miteinander zu vereinbaren. Jedoch ist eine korrekte Anonymisierung, die jeglichen Personenbezug entfernt, kaum zu erreichen und schwerlich zu garantieren, wenn gleichzeitig möglichst viel des Informationsgehalts der Daten erhalten werden soll. Wir geben in diesem Aufsatz einen Überblick über den Stand der Technik der Anonymisierung für strukturierte und unstrukturierte Daten, arbeiten die bestehenden Defizite heraus und formulieren Herausforderungen, die auf dem Weg zu besseren Anonymisierungsverfahren gelöst werden müssen.

Keywords: Anonymisierung; personenbezogene Daten; strukturierte Daten; unstrukturierte Daten; Textdaten; maschinelles Lernen; Informationsverlust

1 Motivation

In der heutigen Zeit werden Daten als das neue Gold oder Öl angesehen. In vielen Szenarien geht es um *Daten über natürliche Personen*, etwa um deren Surfverhalten im Internet, um deren Konsumverhalten, um Bewegungsprofile von Personen, um deren finanzielle oder gesundheitliche Situation und Historie, um die Interessen, Gesinnungen und Kontakte von Personen oder um öffentliche oder private Kommunikation. Menschen haben jedoch generell ein Bedürfnis nach *Privatsphäre* und auch ein Grundrecht (Artikel 7 der EU-Grundrechtecharta) darauf. In Bezug auf Datenverarbeitung wird das Recht auf Privatsphäre durch das Grundrecht auf *Datenschutz* (Artikel 8 der EU-Grundrechtecharta) ergänzt, welches durch die Datenschutzgrundverordnung (DSGVO) konkretisiert ist.

Auf der einen Seite gibt es also das Interesse, viele Daten zu sammeln, zusammenzuführen und auszuwerten. Unter dem Leitmotiv *Big Data* sind die technischen Möglichkeiten und die Erwartungen diesbezüglich stark gestiegen. Insbesondere Verfahren des maschinellen Lernens versprechen gesteigerte Möglichkeiten für Rückschlüsse aus Daten. Auf der anderen Seite gibt es die *betroffenen Personen* mit den Grundrechten auf Privatsphäre und Datenschutz. Diese Personen sind verschiedenen Risiken durch die Datenverarbeitung ausgesetzt, beispielsweise, dass sie durch das Gefühl ständiger Beobachtung ihre Handlungsfreiheit

* Diese Arbeit wurde vom Hessischen Ministerium des Innern und für Sport im Teilprojekt „Privacy und Big Data“ im Rahmen des Projekts „Cybersicherheit für die digitale Verwaltung“ gefördert.

[1] Fraunhofer-Institut für Sichere Informationstechnologie SIT, Rheinstraße 75, 64295 Darmstadt
{vorname}.{nachname}@sit.fraunhofer.de

doi:10.18420/inf2019_53

einschränken, dass sie durch automatisierte Entscheidungen diskriminiert werden oder dass sie durch Fehler in Daten und Systemen und den daraus resultierenden Fehlentscheidungen signifikant benachteiligt werden.

Der Datenschutz dient dazu, die Risiken für Betroffene zu minimieren und rechtliche Mittel zu schaffen, um das Machtverhältnis zwischen Datennutzern und Betroffenen auszugleichen. Von vielen Datennutzern wird der Datenschutz jedoch als Hindernis wahrgenommen, welches möglicherweise sogar manche Vorhaben zur Gewinnung von operativen oder grundsätzlichen Erkenntnissen oder zur Entwicklung und Erprobung neuer Verfahren vereitelt. Dieser Interessenskonflikt muss von Einzelfall zu Einzelfall bewertet und gelöst werden. Generell kann man jedoch unterscheiden, ob eine Datenverarbeitung auf konkrete Personen abzielt, etwa im Endkundengeschäft eines Unternehmens oder bei der Auslieferung von Werbung, oder ob es nur um Erkenntnisse über größere Personengruppen geht, z. B. statistische Eigenschaften, Zusammenhänge und Tendenzen, etwa in der Geschäftsanalytik, in Testumgebungen oder in der Forschung. In der zweiten Konstellation ist die *Anonymisierung* von Daten ein probates Mittel zur Ermöglichung der Datennutzung, da anonymisierte Daten keinen Personenbezug mehr enthalten und somit nicht mehr dem Datenschutz unterliegen.

In den nachfolgenden Abschnitten werden die bestehenden Herausforderungen für die Anonymisierung detailliert herausgearbeitet. Dabei werden primär technische Herausforderungen betrachtet, aber in Abschnitt 2 wird deutlich, dass es auch bei den rechtlichen Rahmenbedingungen Herausforderungen gibt. Technische Anonymisierungsverfahren müssen nach der Art der vorliegenden Daten gewählt werden. Wegen der langen Tradition der Verarbeitung und Anonymisierung von strukturierten Daten untersuchen wir zunächst dieses Gebiet (s. Abschnitt 3). Da viele Information jedoch nicht strukturiert, sondern als Fließtexte vorliegen und zunehmend auch bei automatischen Analysen einbezogen werden, betrachten wir auch Textdaten (s. Abschnitt 4). Aufgrund der zunehmenden Relevanz von maschinellem Lernen gehen wir auch auf die hierdurch begründeten besonderen Herausforderungen für die Anonymisierung ein (s. Abschnitt 5). Nicht genauer betrachtet wird in dieser Publikation die Anonymisierung von Multimediadaten (Bild, Audio, Video) aufgrund der Weite dieses Feldes, welches eine eigenständige Arbeit erfordern würde.

2 Definition und Überprüfbarkeit von Personenbezug und Anonymität

Der Datenschutz regelt den Umgang mit *personenbezogenen Daten*. Grundsätzlich besteht jedoch eine große Schwierigkeit in einer exakten Definition solcher Daten. Analog dazu ist es schwierig zu definieren, wann Daten *nicht personenbezogen*, also *anonym*, sind.

Die DSGVO definiert personenbezogene Daten als solche, „die sich auf eine identifizierte oder identifizierbare natürliche Person [. . .] beziehen“ (Artikel 4 Nr. 1 DSGVO). Während die Identifizierbarkeit von Personen weiter erläutert wird, bleibt unspezifiziert, wie konkret das Beziehen auf eine natürlichen Person sein muss oder wie vage es sein kann, damit ein Personenbezug im Sinne des Gesetzes gegeben ist. Etwas spezifischer in diesem Aspekt

ist die alte Fassung des Bundesdatenschutzgesetzes (BDSGaF), welche personenbezogene Daten als „Einzelangaben“ zu natürlichen Personen definiert (§ 3 Abs. 1 BDSGaF). Hier wird deutlich, dass es nicht um allgemeine statistische Aussagen über Personen geht, sondern um Angaben über einzelne, konkrete Personen. Es ist jedoch zu berücksichtigen, dass in vielen Fällen auch bei Mehrpersonenangaben *Rückschlüsse* über einzelne der einbezogenen Personen getroffen werden können. Dadurch ist der Übergang zwischen Einzelangaben und anonymen Statistiken fließend und es ist nicht klar, wo die Grenze im Sinne des BDSGaF liegt und noch weniger, wo sie im Sinne der DSGVO liegt.

Da in der Datenschutzrichtlinie, welche durch die DSGVO abgelöst wurde, die Definition von personenbezogenen Daten im Kern identisch mit der Definition aus der DSGVO ist, sind die Deutungen der ehemaligen Artikel-29-Datenschutzgruppe zu dem Begriff der personenbezogenen Daten auch im Kontext der DSGVO relevant. In der Opinion 05/2014 [DSG14] befasst sich die Gruppe mit dem Thema Anonymisierung und hier werden Rückschlüsse als eines der zentralen Risiken betrachtet. Dieses Risiko wird folgendermaßen charakterisiert: „Inference, which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes“, wonach eine sehr weitreichende Definition von Rückschlüssen gewählt wird. Demnach kann der Personenbezug von Daten auch weit in allgemeine statistische Aussagen hinein bestehen. Der Interpretationsspielraum beim Begriff des Personen*bezugs* muss ausgeräumt werden, um das Problem der klaren Abgrenzung von personenbezogenen und anonymen Daten zu lösen. Hier ist insbesondere die Rechtsentwicklung gefragt.

Neben einer präzisen, formalen Definition von personenbezogenen Daten fehlt es auch an *Kriterien*, mit denen Daten zweifelsfrei überprüft werden können, ob ein Personenbezug vorhanden ist oder ob die Daten anonym sind. Mangels einer solchen Überprüfbarkeit gibt es keine *Garantie*, dass ein nach dem Stand der Technik anonymisierter Datenbestand auch tatsächlich anonym ist. Viele der bestehenden Anonymisierungsverfahren für strukturierte Daten (s. Abschnitt 3) arbeiten zwar mit gewissen Anonymitätsmaßen als formale Kriterien, aber diese Maße geben eher einen Grad von Anonymität unter bestimmten, teils impliziten, Annahmen über die Daten und über mögliche Angriffe auf die Anonymität wieder. Hier stellt sich zum einen die Frage, welcher Anonymitätsgrad nach dem jeweiligen Maß ausreichend ist, um von Anonymität im Sinne eines nicht mehr vorhandenen Personenbezugs sprechen zu können, und zum anderen, ob nicht auch jenseits der von dem verwendeten Maß betrachteten Aspekte Angriffe auf die Anonymität möglich sind. Für unstrukturierte Daten fehlen Anonymitätskriterien sogar gänzlich.

3 Anonymisierung strukturierter Daten

Strukturierte Daten werden oft tabellarisch dargestellt, so dass jede Spalte ein bestimmtes Attribut enthält und jede Zeile einen einzelnen Datensatz. Bei personenbezogenen Tabellen ist typischerweise je eine Zeile einer Person zugeordnet. Für die Anonymisierung solcher Daten gibt es verschiedene elementare *Strategien*:

Generalisierung: Die jeweiligen Attributwerte werden durch weniger genaue Angaben ersetzt, etwa durch Intervalle bei numerischen Daten oder durch übergeordnete Kategorien bei kategorischen Daten.

Löschung: Der Inhalt einzelner Zellen, Spalten oder Zeilen wird gelöscht. Dies entspricht einer Generalisierung zu einem allumfassenden und nichtssagenden Wert, etwa „*“.

Mikroaggregierung: Die Daten werden nach Ähnlichkeit in den Attributwerten gruppiert (engl. *clustering*) und pro Gruppe werden die einzelnen Werte zu einem repräsentative Werte zusammengefasst, etwa dem Mittelwert oder Median.

Verfälschung: Ein Teil der Daten oder alle Daten werden zufällig abgewandelt. Dies kann z. B. dadurch erreicht werden, dass zu den Werten zufällige Störungen hinzugefügt werden, dass verschiedene Einträge in der Tabelle vertauscht werden oder dass eine künstliche Tabelle unter Orientierung an der Originaltabelle synthetisiert wird.

Neben diesen verschiedenen Ansätzen zur Anonymisierung der Daten gibt es auch die Strategie, nicht die Daten in anonymisierter Form herauszugeben, sondern die gewünschte Analyse zu den Originaldaten in geschützter Umgebung zu bringen, die Analyse dort durchzuführen und nur die Ergebnisse vor der Herausgabe zu anonymisieren. Dies kann einfacher sein und präzisere Ergebnisse liefern, aber es muss stets die rechtliche Zulässigkeit einer solchen Verarbeitung geprüft werden. Zudem schwindet der Vorteil und kann sich in das Gegenteil kehren, wenn viele Analysen durchgeführt werden sollen, da bei der Anonymisierung der Ergebnisse dann auch die Querbeziehungen zwischen allen Ergebnissen berücksichtigt werden müssen.

Zum Bestimmen des Anonymitätsgrades von Daten, die mit den oben genannten Strategien behandelt worden sind, gibt es verschiedene *Kriterien* bzw. *Maße*. Diese Maße unterscheiden sich darin, welche Annahmen über das Hintergrundwissen eines Angreifers und über die Art des zu erreichenden Schutzes gemacht werden. Minimalen Schutz bieten die Kriterien *k-Map* [Sw01] und *δ-Presence* [NAC07], da hier angenommen wird, dass die in der Tabelle erfassten Individuen aus einer größeren Population stammen, ein Angreifer aber nicht wissen kann, ob eine bestimmte Person in der Tabelle enthalten ist. Das bekanntest Anonymitätskriterium ist *k-Anonymität* [Sw01]. Nach diesem Kriterium muss es jeweils mindestens k für eine Person in Frage kommende Einträge in der Tabelle geben, so dass eine Re-Identifikation nicht möglich ist. Da dennoch möglicherweise einzelne Attribute einer Person durch eine k-anonyme Tabelle offengelegt werden können, wurde das Kriterium zu *l-Diversität* [Ma06] und *t-Closeness* [LLV07] weiterentwickelt. Grundsätzlich anders ist das Konzept von *Differential Privacy* [Dw06a]. Hier wird die Anonymität daran gemessen, wie sehr sich das Ergebnis durch Weglassen oder Hinzufügen einer Person ändern kann, und somit wie viel an Information maximal über eine Person offenbart wird.

3.1 Algorithmen für k-Anonymität und verwandte Kriterien

Es gibt eine Vielzahl von *Algorithmen* zum Erreichen der verschiedenen Anonymitätskriterien. Insbesondere für k-Anonymität und die daran anknüpfenden Maße gibt es eine Vielzahl von Algorithmen basierend auf Generalisierung und Löschung. Einfachere Algorithmen beschränken sich auf eine Generalisierung auf der Attribut-Ebene, d. h. es wird für eine Tabellenspalte insgesamt festgelegt, welcher Wert zu welchem generalisiert wird („global recoding"), während komplexere Algorithmen die Generalisierung auf Zell-Ebene festlegen können („local recoding"). Die zweite Gruppe von Algorithmen kann das Ziel mit weniger Informationsverlust erreichen, jedoch ist die Durchführung bei realen Tabellen meist zu aufwändig, da der Aufwand zum Finden einer optimalen Generalisierung bei naiver Suche exponentiell mit der Anzahl der Tabellenzellen steigt. Diese Optimierungsaufgabe ist in der Tat NP-schwer [Du07], so dass keine effizienten Algorithmen existieren. Aber selbst die erste Gruppe von Algorithmen kann bei großen Datentabellen, insbesondere, wenn viele Attribute vorhanden sind, zu aufwändig werden. Algorithmen auf Basis von Mikroaggregierung können eine gute Effizienz aufweisen und gleichzeitig mehr Informationen erhalten als Generalisierungen auf Attribut-Ebene.

Bei allen Verfahren, die k-Anonymität oder verwandte Eigenschaften auf den Daten sicherstellen, ist zu beachten, dass diese Eigenschaften keine Anonymität garantieren (vgl. Abschnitt 2). Sie schützen nur gegen bestimmte Risiken und auch nur, wenn die Annahmen über das Hintergrundwissen der Angreifer und über die Eigenschaften der Daten korrekt sind. Um das Problem von nicht berücksichtigten Angriffsmöglichkeiten zu lösen, muss die Forschung entweder ultimative Anonymitätskriterien finden oder sie muss wenigstens Anwender dabei unterstützen, Schutzlücken oder unpassende Annahmen aufzudecken. Für zweiteres sollten die existierenden Anonymitätskriterien durch formale *Angreifermodelle* ergänzt werden, welche die angenommenen Fähigkeiten und das angenommene (Hintergrund-) Wissen von Angreifern explizit machen. Anwender können damit leichter erkennen, was in ihren Szenarien durch welche Anonymisierung tatsächlich geschützt wird. Zusätzlich muss der Anwender aber auch stets die Semantik der vorhanden Datenattribute bei der Wahl der Anonymisierung beachten.

3.2 Algorithmen für Differential Privacy

Für Differential Privacy gibt es ebenfalls eine Reihe von Algorithmen. Diese Algorithmen nutzen die Strategie der zufälligen Verfälschung. Der *Laplace-Mechanismus* [Dw06b] ist für Frage-Antwort-Systeme geeignet, bei denen nur die aus den Originaldaten gewonnen Antworten in anonymisierter Form herausgegeben werden sollen. Dazu gibt es ein Privacy-Budget, welches nach und nach von den Antworten aufgebraucht wird, so dass sukzessive der Informationsgehalt von Antworten reduziert (d. h. die Störung erhöht) wird oder irgendwann gar keine Antworten mehr gegeben werden können, wenn das Privacy-Budget verbraucht ist. Der *Exponential-Mechanismus* [MT07] hingegen kann genutzt werden, um synthetische

Tabellen nach dem Vorbild der Originaltabelle zu erzeugen [BLR08]. Dabei wird zum einen die Nähe zu den Originaldaten mit höheren Wahrscheinlichkeiten begünstigt und zum anderen wird über einen Parameter gesteuert, wie groß die Streuung ist, um das Kriterium von Differential Privacy zu erfüllen. Der Exponential-Mechanismus ist jedoch mit sehr hohem Aufwand verbunden.

Ein weiteres Verfahren für Differential Privacy sind *randomisierte Antworten*, welche bereits lange in sozialwissenschaftlichen Studien eingesetzt werden [Wa65]. Dabei geben Probanden in Abhängigkeit von Münzwürfen oder Ähnlichem zufallsbestimmte oder wahrheitsgemäße Antworten. So lässt sich aus der einzelnen Antwort keine Wahrheit ablesen, d. h. die Privatsphäre der Probanden wird schon bei der Datenerhebung geschützt. Durch das Gesetz der großen Zahlen kann der Einfluss der Zufallsantworten auf die Gesamtheit der Antworten näherungsweise herausgerechnet werden, so dass mit statistischen Methoden Erkenntnisse aus den Daten abgeleitet werden können. Die Anforderungen von Differential Privacy werden bei geeignetem Studienaufbau in der Tat erfüllt [Ka08]. Dadurch ist ein effektiver Schutz der Privatsphäre gegeben und zudem ist das Verfahren rechentechnisch (aber nicht unbedingt für die Probanden) effizient. Zu beachten bleibt aber, dass ein deutlicher Informationsverlust entsteht und dass die Daten in ihren statistischen Eigenschaften hochgradig verändert werden. Letzteres kann rechnerisch korrigiert werden, aber ersteres kann nur mit einer größeren Probandenzahl kompensiert werden.

4 Anonymisierung von Texten

Bei Textdokumenten unterscheiden wir zwischen der Metadatenebene, der Inhaltsebene und der Schreibstilebene. Auf all diesen Ebenen können Personenbezüge vorhanden sein. Bevor wir auf die Anonymisierung hinsichtlich jeder einzelnen Ebene eingehen, klären wir zunächst diese Begriffe. Die *Metadatenebene* ist eine vom Text entkoppelte Ebene, die Zusatzinformationen zu einem Dokument bereitstellt. Die *Inhaltsebene* ist die zentrale Ebene, die die eigentliche Information trägt. Die *Schreibstilebene* ist in die Inhaltsebene eingebettet und lässt sich nicht ohne Weiteres von dieser entkoppeln.

4.1 Anonymisierung auf der Metadatenebene

Die Existenz und Form von Metadaten hängt davon ab, in welchem Format ein Dokument vorliegt. Handelt es sich um eine Datei in einem komplexen Format (z. B. eine PDF-Datei oder ein Word-Dokument), so liegen in der Regel Metadaten vor. Diese enthalten Felder wie etwa Autoren, Titel, Schlüsselwörter und Erstellungsdatum und reichern das Dokument mit semantischen Informationen an. Handelt es sich jedoch um eine reine Textdatei, so existiert innerhalb der Datei keine Metadatenebene. Gegebenenfalls finden sich jedoch Metadaten im umgebenden System, welches die Datei speichert, was beispielsweise ein Dateisystem oder eine E-Mail sein kann.

Metadaten bergen die Gefahr, dass sie oft vom Ersteller nicht wahrgenommen werden, jedoch Informationen enthalten, die dessen Identität ungewollt preisgeben können. Die Anonymisierung der Metadatenebene ist meist trivial durchführbar, indem die Metadaten entweder gar nicht erst erstellt oder nachträglich entfernt werden.

4.2 Anonymisierung auf der Inhaltsebene

Inhaltsdaten enthalten oftmals Entitäten wie z. B. Personennamen, Bezeichnungen von Firmen oder Organisationen oder geographische Orte, die die Identität des Autors oder die von Dritten referenzieren können. Diese lassen sich anders als Metadaten nicht mit einfachen Mitteln entfernen,[2] ohne die Semantik des Dokuments zu verletzen. Die Voraussetzung für die Anonymisierung von Texten ist, zunächst die Verweise auf Identitäten zu identifizieren. Diese können mithilfe computerlinguistischer Verfahren wie *Eigennamenerkennung* (engl. *named entity recognition*) ermittelt werden [Li18a; YB18]. Anschließend können diese Verweise mit verschiedenen Strategien anonymisiert werden. Eine hundertprozentige Erkennung aller Verweise ist jedoch nicht möglich, sodass immer ein Restrisiko verbleibt.

Eine Möglichkeit zur Anonymisierung entsprechender Textstellen läuft über eine Pseudonymisierung mittels partieller Verschlüsselung. Dabei werden Verweise auf Identitäten mit einem geheimen Schlüssels k verschlüsselt, sodass aus dem Dokument $\mathcal{D}$ ein modifiziertes Dokument $\mathcal{D}'$ entsteht. $\mathcal{D}'$ kann somit nur von autorisierten Personen, die k besitzen, entschlüsselt und dadurch vollständig gelesen werden. Stellt man sicher, dass nach der Pseudonymisierung niemand mehr den Schlüssel k hat, ist eine Anonymisierung erreicht. Der Nachteil der partiellen Verschlüsselung ist, dass der Lesefluss in $\mathcal{D}'$ durch die verschlüsselten Elemente gestört wird und das Dokument daher nur fragmentarisch gelesen werden kann, was den Nutzen des Dokuments reduziert.

Eine Alternative zur partiellen Verschlüsselung ist, die Verweise auf Identitäten zu *paraphrasieren*. Damit kann eine Anonymisierung erreicht und gleichzeitig die Semantik von $\mathcal{D}$ bis zu einem gewissen Grad beibehalten werden. Analog zur partiellen Verschlüsselung führt dies zwar ebenfalls zu einem Informationsverlust, allerdings in einer Form, bei der zum einen die modifizierte Version $\mathcal{D}'$ vollständig lesbar bleibt und zum anderen niemand mit einer Art Schlüssel die Ursprungsinformation wiederherstellen kann. Dazu gilt es, die identifizierten Entitäten durch generischere Angaben[3] zu ersetzen.

Eine wichtige Frage bei der Paraphrasierung ist, woher die abgewandelten Entitäten bezogen werden können. Eine Möglichkeit besteht darin, vorhandene linguistische Ressourcen zu verwenden wie etwa *Ontologien* oder *lexikalische Wortnetze*, mit denen semantisch sinnvolle Ersetzungen durchgeführt werden können. Diese müssen in der Regel händisch erstellt werden und sind dadurch mit entsprechenden Aufwand und hohen Kosten verbunden. Hinzu

[2] Ausgenommen sind isolierte Entitäten, die unabhängig vom Text sind (z. B. der Name nach einer Grußformel).
[3] Beispielsweise „Angela Merkel“ → { „deutsche Politikerin“, „gebürtige Hamburgerin“, . . . }.

kommt die Problematik der temporalen Veränderung von Sprachen,[4] sodass gegebenenfalls zu einer Entität x in einem Text keine passenden Ersetzungen in einer Wortliste gefunden werden können, da die Wortliste zu einem Zeitpunkt erstellt wurde, als x noch nicht existierte. Alternativ zu händisch erstellten linguistischen Ressourcen eignen sich Ansätze basierend auf sogenannte *Word Embeddings*. Die Idee dahinter ist, Wörter eines Vokabulars als reelle Vektoren in einem hochdimensionalen Raum darzustellen und diesen auf einen Raum mit niedrigerer Dimension abzubilden, sodass im zweiten Raum semantische Beziehungen der Wörter durch die Nähe der entsprechenden Vektoren widergespiegelt werden. Mithilfe solcher Word Embeddings lassen sich ohne den Einsatz gelabelter Daten bzgl. einer Entität x semantisch ähnliche Entitäten $y_1, y_2, \ldots$ finden, die eine Ersetzung erlauben. Vorausgesetzt werden hier jedoch genügend ungelabelte Textdaten, welche Informationen über die Entität x enthalten. Ein wesentlicher Nachteil hierbei ist allerdings, dass die Entitäten nicht in einer festgelegten Relation (z. B. Synonymie) zueinander stehen, sondern sich über mehrere Relationen wie etwa Hyperonymie, Hyponymie, Meronymie oder Holonymie erstrecken können. Der Literatur zufolge existiert noch kein zufriedenstellender Ansatz mit dessen Hilfe Entitäten hinsichtlich ihrer semantischen Relationen automatisiert abgegrenzt werden können, sodass es hierfür noch weitere Forschungsarbeit bedarf.

4.3 Anonymisierung auf der Schreibstilebene

Die Identität einer Person lässt sich auch über dessen Schreibstil bestimmen. Im Laufe des letzten Jahrzehnts hat sich die *digitale Textforensik* als Forschungsfeld etabliert. Hauptaugenmerk liegt dabei auf der *Autorschaftsanalyse*, welche das Ziel verfolgt, Informationen über die Autoren digitaler Dokumente offenzulegen [Po19].

Aus der Notwendigkeit heraus, die Identität von Autoren zu schützen, entstand das Forschungsfeld *Author Obfuscation* (AO), welches sich damit befasst, wie sich der Schreibstil in Dokumenten verschleiern lässt. Bisherige AO-Ansätze lassen sich in manuelle, computerassistierte und automatische Verfahren aufteilen [GA19], wobei der Forschungsfokus insbesondere auf letzteren liegt. Automatische AO gilt als sehr anspruchsvoll, da sie auf Sprachkompetenzen zurückgreifen muss, um anonymisierende Umformungen in den Dokumenten vorzunehmen unter gleichzeitiger Beibehaltung der ursprünglichen Semantik.

Unter den veröffentlichten automatischen AO-Verfahren ist vor allem der Ansatz *Adversarial Author Attribute Anonymity Neural Translation* (A^4NT) von Shetty et al. [SSF18] hervorzuheben. Das Verfahren ist unserer Recherche nach das einzige Verfahren, das eine dedizierte Komponente für die Semantikerhaltung enthält. A^4NT verfolgt eine intuitive Idee, die analog zu einer maschinellen Übersetzung funktioniert. Während in der maschinellen Übersetzung ein Dokument in eine festgelegte Zielsprache übersetzt wird, wird bei A^4NT das Dokument in dieselbe Sprache wie die Quellsprache „übersetzt“, um den Schreibstil des ursprünglichen Autors nicht mehr wiedererkennen zu können. Das Verfahren wurde

[4] Vor 20 Jahren gab es z. B. noch nicht die Wörter „googlen“, „Podcast“ und „Smombie“.

hinsichtlich der drei autorspezifischen Attribute Alter (unter 20 vs. über 20), Geschlecht und Identität (Obama vs. Trump) anhand einer Kollektion von Blogartikeln und einer Kollektion von politischen Reden getestet. Hinsichtlich der Attribute Alter und Geschlecht konnten Shetty et al. die Erkennungsgenauigkeit (F_1-Wert) beim Alter von 88 % auf 8 %, beim Geschlecht von 75 % auf 39 % und bei der Identität von 100 % auf 0 % senken, was dafür spricht, dass eine Anonymisierung auf der Schreibstilebene möglich ist.

5 Anonymitätsrisiken beim maschinellen Lernen

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 – die entscheidenden Verknüpfungen gefunden werden können, welche das Herstellen von Personenbezügen ermöglichen. 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. Bestimmte ML-Techniken können sich jedoch unerwartet deutlich an die zum Training des Modells verwendeten Daten erinnern. So speichern Support Vector Machines 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 [AC19].

Fredrikson et al. [FJR15] demonstrierten, dass die Erinnerung in neuronalen Netzen, welche zur Gesichtserkennung genutzt wurden, mitunter so stark sein kann, dass es möglich ist, ein Abbild der Trainingsdaten zu rekonstruieren – ein sogenannter *Modellinversionsangriff*. Shokri et al. [Sh17] bewiesen, dass neuronale Netze aufgrund ihrer Konstruktion anfällig für *Membership-Inference-Angriffe* sind. Die Autoren wiesen nach, dass ein trainiertes Netz merkbar anders auf Informationen reagiert, welche bereits zum Training verwendet wurden als auf bisher ungesehene Testdaten. Aufgrund dieser Rückmeldung kann ein Angreifer eindeutig zuordnen, ob ein Individuum in einem bestimmten Datensatz enthalten ist oder nicht. Solche Angriffe stellen allgemein eine Verletzung der Privatheit dar, sind aber besonders dann kritisch, wenn es sich um sensible Informationen handelt, wie beispielsweise Insolvenz oder ob eine bestimmte Krankheit vorliegt.

Das Forschungsfeld *Privacy Preserving Machine Learning* (PPML) ist noch recht jung. Auch wenn es bereits vielversprechende Ansätze gibt, besteht noch viel Entwicklungsbedarf. Nachfolgend werden die wichtigsten Forschungsrichtungen auf diesem Gebiet skizziert.

5.1 Kollaboratives maschinelles Lernen

Würden alle oder viele Personen ihre Daten nicht mehr für Forschungs- und Auswertungszwecke zur Verfügung stellen, hätte das einschneidende Konsequenzen für die Forschung,

insbesondere für die Medizinforschung. Außerdem könnten viele weitverbreitete, nützliche Dienste nicht weiter angeboten werden. Ziel ist es folglich, Daten auf privatsphärenfreundliche Weise einem ML-System zur Verfügung stellen zu können.

Kryptographische Verfahren für kollaboratives Lernen Ein vielversprechender Ansatz, um die Privatheit des Einzelnen zu schützen und gleichzeitig das Training von Modellen auf Daten von vielen Personen zu ermöglichen, ist die *sichere Mehrparteienberechnung* (engl. *secure multi-party computation*, MPC) als Teilgebiet der Kryptographie. Das Ziel von MPC ist 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 und deren Abbruchwahrscheinlichkeit existieren verschiedene Ansätze mit unterschiedlichem Rechen- und Kommunikationsaufwand, um dieses Ziel zu erreichen. Tatsächlich gibt es erste MPC-Ansätze im Kontext von maschinellem Lernen zur Summenberechnung von Modellparametern [Bo17].

Homomorphe Verschlüsselung erlaubt – im Gegensatz zu herkömmlichen Verschlüsselungsmethoden – Rechenoperationen direkt auf den verschlüsselten Daten auszuführen, ohne diese zuvor in Klartext zu überführen 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. Mit homomorpher Verschlüsselung können daher Daten an eine nicht-vertrauenswürdige Instanz weitergegeben werden und Berechnungen dort durchgeführt werden. Insbesondere die sogenannte voll-homomorphe Verschlüsselung generiert jedoch einen signifikanten Rechenmehraufwand [Do16; Li18b], der diese für rechenintensive Anwendungen wie maschinelles Lernen bisher unbrauchbar macht. Erste praktikable Ansätze verwenden daher Vereinfachungen. So wenden Dowlin et al. [Do16] ein auf unverschlüsselten Rohdaten trainiertes neuronales Netz auf „somewhat“ homomorph verschlüsselte Daten an. Long et al. [Lo18] haben das Training verschiedener ML-Verfahren mit additiv-homomorpher Verschlüsselung und Zero-Knowledge-Beweisen realisiert.

Dezentrales maschinelles Lernen Eine Lösung zum privatsphärenfreundlichen Lernen auf Daten von vielen Nutzern ist das dezentrale Lernen. Hierbei trainieren die Nutzer ein Grundmodell lokal auf ihren individuellen Daten und übermitteln lediglich die neu berechneten Gradienten des Trainings oder die neuen Modellparameter 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 oder Parameter zurück an den Server [Mc17]. Hauptsächlich zum Schutz der Privatsphäre, aber auch zur Kommunikationseffizienz, erlaubt der

Ansatz nach Shokri und Shmatikov [SS15], dass nicht alle Aktualisierungen mit dem Server geteilt werden müssen, sondern nur eine kleine Teilmenge, deren Größe vom Nutzer selbst festgelegt wird. Allerdings sollte sich der Nutzer des Trade-offs zwischen der Menge der geteilten Aktualisierungen sowie Trainingszeit und -qualität bewusst sein.

Hitaj et al. [HAP17] haben nachgewiesen, dass es selbst in solchen dezentralen Lernansätzen mit Hilfe eines Generative Adversarial Networks (GAN) möglich ist, über die übrigen aufrichtigen Teilnehmer sensible Daten zu sammeln. Melis et al. [Me19] entkräften teilweise die Angriffe von Hitaj et al., zeigen aber selbst neue Angriffsstrategien auf.

5.2 Differential Privacy für maschinelles Lernen

Arbeiten zu *Differentially Privacy* im Kontext des maschinellen Lernens (vgl. Differential Privacy in Abschnitt 3) erforschen verschiedene Aspekte des Verrauschens von potentiell angreifbaren Daten. Untersucht wird hier meist, auf welcher Ebene die Störungen idealerweise Eingang in den Algorithmus finden – ob nun auf Input- oder Output-Ebene oder ob die Gradienten oder die Verlustfunktion verrauscht werden – und welche Verteilungseigenschaften das Rauschen selbst haben sollte. Das Ziel ist, einen optimalen Trade-off zwischen Privatheit und Ergebnisqualität zu erreichen.

Eine andere Richtung verfolgt der Ansatz der *Differentially Private Data Synthesis* (DIPS). Hierbei werden Daten auf Basis realer Datensätze beispielsweise mittels Copula-Funktionen [LXJ14] oder Generative Adversarial Networks [TF19] 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 anderen Datensyntheseverfahren. Darüber hinaus besitzen die Daten annähernd die gleichen Verteilungseigenschaften wie die zugrundeliegenden Originaldaten und können in beliebiger Anzahl generiert werden, um so beispielsweise die Güte eines ML-Models zu verbessern [ML19; PCN18].

5.3 Generelle Limitationen der bestehenden Verfahren

Angriffspunkt für die weitere Forschung ist u. a. die Anwendbarkeit der Verfahren bzw. deren mangelnde Flexibilität. Die meisten privatheiterhaltenden Verfahren sind nur für die Anwendung auf einen bestimmten Lernalgorithmus optimiert und auf andere ML-Verfahren schwer bis gar nicht übertragbar. Zudem stellt mangelnde Skalierbarkeit ein Hindernis für die Anwendung privatheiterhaltender Maßnahmen in der Praxis dar. Das Schützen sensibler Informationen generiert immer zusätzliche Kosten – entweder 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 groß aus, dass eine Anwendung in der Praxis nicht tragbar ist [AC19].

6 Zusammenfassung der Herausforderungen und Fazit

Grundsätzlich ist eine exakte Definition von Personenbezug und Anonymität nötig, an der die rechtliche Einordnung von Daten zweifelsfrei entschieden werden kann und an der Anonymitätskriterien zur praktischen Prüfung von Daten gemessen werden können. Zudem ist eine Weiterentwicklung auf dem Gebiet der Anonymitätskriterien nötig. Zum einen existieren solche Kriterien hauptsächlich für tabellarische Daten. Zum anderen mangelt es den meisten dieser Kriterien an starken Garantien, so dass etwa Angreifer mit zusätzlichem Hintergrundwissen weitere Informationen über konkrete Personen extrahieren können. Daher müssen diese Kriterien durch eine theoretische Untersuchung basierend auf zu entwickelnden formalen Angreifermodellen hinsichtlich ihrer Garantien präzisiert werden.

Für strukturierte Daten sind die heutigen Kernmethoden zur Anonymisierung hauptsächlich vor zehn bis zwanzig Jahren publiziert worden und es wurden bereits viele Verbesserungen entwickelt. Handlungsbedarf besteht aber weiterhin neben den bereits genannten Problemen der Anonymitätskriterien auch in Bezug auf die Minimierung des Informationsverlustes bei gleichzeitiger Maximierung der Effizienz von Algorithmen, insbesondere in Bezug auf eine adaptive Ausbalancierung dieser entgegengesetzten Anforderungen.

Bei der Anonymisierung der Inhaltsebene von Textdaten gibt es nach wie vor die Herausforderung der zuverlässigen Erkennung von Entitäten sowie Herausforderungen in Bezug auf Umsetzbarkeit und Anwendbarkeit von Strategien zur Ersetzung dieser Entitäten. In Bezug auf die Anonymisierung der Stilebene gibt es erste empirische Ergebnisse, die erfolgversprechend sind, aber eine allgemeine Zuverlässigkeit kann noch nicht daraus geschlossen werden. Zudem ist eine Anonymitätsgarantie bei Texten noch weniger möglich als bei strukturierten Daten.

Der Privatsphärenschutz in Verbindung mit maschinellem Lernen ist ein noch recht junges und unerforschtes Thema, das erst vor wenigen Jahren verschiedene Risiken aufgedeckt hat. Erste Lösungsansätze, etwa in Verbindung mit Differential Privacy oder Kryptographie, beschränken sich hauptsächlich auf den Schutz additiver Operationen. Neben den allgemeinen Herausforderungen dieser Schutzstrategien sind hier auch die Herausforderungen der Anwendbarkeit und Effektivität im Kontext des maschinellen Lernens zu lösen.

Diese Arbeit hat gezeigt, dass es bereits viele wissenschaftliche Publikation zur Anonymisierung von Daten gibt. Die Literatur behandelt sowohl Anonymisierungskonzepte als auch Limitationen und Schutzlücken der Konzepte. Bei strukturierten Daten ist unter Berücksichtigung der Einschränkungen in Bezug auf Anonymitätsgarantien und Algorithmeneffizienz ein Praxiseinsatz von Anonymisierung bereits möglich, während auf den anderen untersuchten Gebieten die Anonymisierungsstrategien hauptsächlich prototypische Forschungsarbeiten sind. In allen Bereichen gibt es noch viele zu lösende Forschungsfragen, die hier herausgearbeitet wurden.

Literatur

[AC19] Al-Rubaie, M.; Chang, J. M.: Privacy-Preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy 17/2, S. 49–58, 2019.

[BLR08] Blum, A.; Ligett, K.; Roth, A.: A Learning Theory Approach to Non-Interactive Database Privacy. In: STOC'08. ACM, S. 609–618, 2008.

[Bo17] Bonawitz, K. et al.: Practical Secure Aggregation for Privacy-Preserving Machine Learning. In: CCS 2017. ACM, S. 1175–1191, 2017.

[Do16] Dowlin, N. et al.: CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy, Techn. Ber., 2016.

[DSG14] Artikel-29-Datenschutzgruppe: Opinion 05/2014 on Anonymisation Techniques, Techn. Ber. WP216, Artikel-29-Datenschutzgruppe, 10. Apr. 2014.

[Du07] Du, Y. et al.: On Multidimensional k-Anonymity with Local Recoding Generalization. In: ICDE 2007. IEEE, 2007.

[Dw06a] Dwork, C.: Differential Privacy. In: Automata, Languages and Programming. Springer, S. 1–12, 2006.

[Dw06b] Dwork, C. et al.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Theory of Cryptography. Springer, S. 265–284, 2006.

[FJR15] Fredrikson, M.; Jha, S.; Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: CCS 2015. ACM, S. 1322–1333, 2015.

[GA19] Gröndahl, T.; Asokan, N.: Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace?, 26. Feb. 2019, arXiv: 1902.08939v2.

[HAP17] Hitaj, B.; Ateniese, G.; Pérez-Cruz, F.: Deep models under the GAN: information leakage from collaborative deep learning. In: CCS 2017. ACM, S. 603–618, 2017.

[Ka08] Kasiviswanathan, S. P. et al.: What Can We Learn Privately? In: FOCS 2008. IEEE Computer Society, S. 531–540, 2008.

[Li18a] Li, J. et al.: A Survey on Deep Learning for Named Entity Recognition, 22. Dez. 2018, arXiv: 1812.09449v1.

[Li18b] Liu, Q. et al.: A survey on security threats and defensive techniques of machine learning: A data driven view. IEEE Access 6/, S. 12103–12117, 2018.

[LLV07] Li, N.; Li, T.; Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: ICDE 2007. IEEE, S. 106–115, 2007.

[Lo18] Long, Y. et al.: Distributed and Secure ML with Self-tallying Multi-party Aggregation, 26. Nov. 2018, arXiv: 1811.10296v1.

[LXJ14] Li, H.; Xiong, L.; Jiang, X.: Differentially private synthesization of multi-dimensional data using copula functions. In: EDBT 2014. OpenProceedings, S. 475–486, 2014.

[Ma06] Machanavajjhala, A. et al.: *l*-Diversity: Privacy Beyond *k*-Anonymity. In: ICDE'06. IEEE, Apr. 2006.

[Mc17] McMahan, H. B. et al.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: AISTATS 2017. Bd. 54. PMLR, S. 1273–1282, 2017.

[Me19] Melis, L. et al.: Exploiting unintended feature leakage in collaborative learning. In: IEEE S&P 2019. 2019.

[ML19] McKay Bowen, C.; Liu, F.: Comparative study of differentially private data synthesis methods, 8. Jan. 2019, arXiv: 1602.01063v4.

[MT07] McSherry, F.; Talwar, K.: Mechanism Design via Differential Privacy. In: FOCS 2007. IEEE Computer Society, S. 94–103, 2007.

[NAC07] Nergiz, M. E.; Atzori, M.; Clifton, C. W.: Hiding the Presence of Individuals from Shared Databases. In: SIGMOD'07. ACM, S. 665–676, 2007.

[PCN18] Page, H.; Cabot, C.; Nissim, K.: Differential privacy an introduction for statistical agencies. In: NSQR. Government Statistical Service, 2018.

[Po19] Potthast, M. et al.: A Decade of Shared Tasks in Digital Text Forensics at PAN. In: Advances in Information Retrieval. Springer, S. 291–300, 2019.

[Sh17] Shokri, R. et al.: Membership inference attacks against machine learning models. In: IEEE S&P 2017. S. 3–18, 2017.

[SS15] Shokri, R.; Shmatikov, V.: Privacy-preserving deep learning. In: ACM CCS 2015. ACM, S. 1310–1321, 2015.

[SSF18] Shetty, R.; Schiele, B.; Fritz, M.: A^{4}NT: Author Attribute Anonymity by Adversarial Training of Neural Machine Translation. In: USENIX Security '18. S. 1633–1650, 2018.

[Sw01] Sweeney, L.: Computational Disclosure Control, A Primer on Data Privacy Protection, Diss., Massachusetts Institute of Technology, Mai 2001.

[TF19] Triastcyn, A.; Faltings, B.: Generating artificial data for private deep learning. In: PAL. Bd. Vol-2335. CEUR Workshop Proceedings, S. 33–40, 18. März 2019.

[Wa65] Warner, S. L.: Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. JASA 60/309, S. 63–69, 1965.

[YB18] Yadav, V.; Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In: COLING 2018. ACL, S. 2145–2158, Aug. 2018.