CHRISTIAN WINTER / VERENA BATTIS / OREN HALVANI

Herausforderungen für die Anonymisierung von Daten

Technische Defizite, konzeptuelle Lücken und rechtliche Fragen bei der Anonymisierung von Daten

Personenbezogene Daten
Strukturierte Daten
Textdaten
Maschinelles Lernen

■ 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 – vor allem, wenn gleichzeitig möglichst viel des Informationsgehalts der Daten erhalten bleiben soll. Der Aufsatz gibt einen Überblick über den Stand der Technik der Anonymisierung für strukturierte und unstrukturierte Daten, arbeitet die bestehenden Defizite heraus und formuliert Herausforderungen, die auf dem Weg zu besseren Anonymisierungsverfahren gelöst werden müssen.

■ Companies, scientists and state agencies have great interest in gaining new insights from data. In so doing, data protection rules must be followed. Anonymization is – at first glance – an attractive solution in order to balance data protection and analysis interests. However, correct anonymization which removes all personal reference can barely be attained and is hard to ensure – in particular, if at the same time as much information should remain intact as possible. This article gives an overview on the status of technology of anonymization for structured and unstructured data, identifies the existing deficits and identifies challenges which need to be solved for a better anonymization process.
Lesedauer: 19 Minuten

I. Motivation

Menschen haben generell ein Bedürfnis nach Privatsphäre und auch ein Grundrecht (Art. 7 GRCh) darauf. In Bezug auf Datenverarbeitung wird das Recht auf Privatsphäre durch das Grundrecht auf Datenschutz (Art. 8 GRCh) ergänzt, welches durch die DS-GVO konkretisiert ist. In vielen Lebensbereichen 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.

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 Kapiteln werden die bestehenden Herausforderungen für die Anonymisierung detailliert herausgearbeitet. Dabei werden primär technische Herausforderungen betrachtet. In Kapitel II. wird jedoch 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 wird zunächst dieses Gebiet untersucht (s. unter III.). Da viele Information jedoch nicht strukturiert, sondern als Fließtexte vorliegen und zunehmend auch bei automatischen Analysen einbezogen werden, betrachten wir auch Textdaten (s. unter IV.). Auf Grund der zunehmenden Relevanz von maschinellem Lernen gehen wir auch auf die hierdurch begründeten besonderen Herausforderungen für die Anonymisierung ein (s. unter V.). Nicht genauer betrachtet wird in dieser Publikation die Anonymisierung von Multimediadaten (Bild, Audio, Video), was auf Grund der Weite dieses Felds eine eigenständige Arbeit erfordern würde.

II. 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 DS-GVO definiert personenbezogene Daten als solche, „die sich auf eine identifizierte oder identifizierbare natürliche Person … beziehen" (Art. 4 Nr. 1 DS-GVO). Während die Identifizierbarkeit von Personen weiter erläutert wird, bleibt unspezifiziert, wie konkret das Beziehen auf eine natürliche Person sein muss, damit ein Personenbezug i.S.d. Gesetzes gegeben ist. Etwas spezifischer in diesem Aspekt ist die alte Fassung des BDSG (BDSG a.F.), welche personenbezogene Daten als „Einzelangaben" zu natürlichen Personen definiert (§ 3 Abs. 1 BDSG a.F.). 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 i.S.d. BDSG a.F. liegt und noch weniger, wo sie i.S.d. DS-GVO liegt.

Da in der Datenschutzrichtlinie (RL 95/46/EG – DS-RL), welche durch die DS-GVO abgelöst wurde, die Definition von personenbezogenen Daten im Kern identisch mit der Definition aus der

DS-GVO ist, sind die Deutungen der ehemaligen *Art. 29-Datenschutzgruppe* zu dem Begriff der personenbezogenen Daten auch im Kontext der DS-GVO relevant. In Opinion 05/2014[1] werden Rückschlüsse als eines der zentralen Risiken bei der Anonymisierung 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 Personenbezugs 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, anhand derer Daten zweifelsfrei daraufhin 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. unter III.) 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 jeweils ausreichend ist, 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.

III. Anonymisierung strukturierter Daten

Strukturierte Daten werden oft tabellarisch dargestellt, sodass 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: 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 repräsentativen Wert 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.

Daneben gibt es auch die Strategie, die gewünschte Analyse auf den Originaldaten in geschützter Umgebung durchzuführen und nur die Ergebnisse vor der Herausgabe zu anonymisieren. Dies kann einfacher sein und präzisere Ergebnisse liefern als in einem ersten Schritt die Datentabelle zu anonymisieren, aber es muss stets die rechtliche Zulässigkeit einer solchen Verarbeitung geprüft werden. Werden jedoch viele Analysen auf einer Datentabelle durchgeführt, so schwindet der Vorteil und kann sich sogar ins Gegenteil kehren, da bei der Anonymisierung der Ergebnisse dann auch die Querbeziehungen zwischen allen Ergebnissen berücksichtigt werden müssen.

Zum Bestimmen des Anonymitätsgrads von Daten, die mit den o.g. 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[2] und delta-Presence[3], 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 bekannteste Maß ist k-Anonymität[4]. Nach diesem Kriterium muss es jeweils mindestens k für eine Person in Frage kommende Einträge in der Tabelle geben. Da einzelne Attribute einer Person durch eine k-anonyme Tabelle offengelegt werden können, wurde das Kriterium zu l-Diversität[5] und t-Closeness[6] weiterentwickelt. Grundsätzlich anders ist das Konzept von Differential Privacy[7]. Hier wird die Anonymität daran gemessen, wie sehr sich das Ergebnis einer Verarbeitung durch In- oder Exklusion einer Person im zu Grunde liegenden Datenbestand ändern kann, und somit wie viel an Information maximal über eine Person offenbart wird.

1. Algorithmen für k-Anonymität und verwandte Kriterien

Für k-Anonymität und die daran anknüpfenden Maße gibt es eine Vielzahl von Algorithmen basierend auf Generalisierung und Löschung. Die einfacheren Algorithmen darunter beschränken sich auf eine Generalisierung auf der Attribut-Ebene, d.h. es wird für eine Tabellenspalte insgesamt festgelegt, welcher Wert zu welchem generalisiert wird (engl. „global recoding"), während komplexere Algorithmen die Generalisierung auf Zell-Ebene festlegen können (engl. „local recoding"). Die zweite Gruppe von Algorithmen kann das Ziel mit weniger Informationsverlust erreichen, jedoch ist die Durchführung bei realen Tabellen meist zu aufwändig. 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 effizient sein 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. unter II.). Sie schützen nur gegen bestimmte Risiken und auch nur, wenn die getroffenen 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 Letzteres sollten die existierenden Anonymitätskriterien durch formale Angreifermodelle ergänzt werden, welche die angenommenen Fähigkeiten und das angenommene (Hintergrund-)Wissen von Angreifern explizit machen.

---

1 *Art. 29-Datenschutzgruppe*, Opinion 05/2014 on Anonymisation Techniques, 2014.
2 *Sweeney*, Computational Disclosure Control: A Primer on Data Privacy Protection, 2001.
3 *Nergiz/Atzori/Clifton*, in: SIGMOD 2007, p. 665.
4 *Sweeney* (o. Fußn. 2).
5 *Machanavajjhala* et al., in: ICDE 2006, Artikel-ID 24.
6 *Li/Li/Venkatasubramanian*, in: ICDE 2007, p. 106.
7 *Dwork*, in: Automata, Languages and Programming, 2006, p. 1.

2. Algorithmen für Differential Privacy

Algorithmen für Differential Privacy nutzen die Strategie der zufälligen Verfälschung. Der Laplace-Mechanismus[8] 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, das nach und nach von den Antworten aufgebraucht wird, sodass irgendwann gar keine Antworten mehr gegeben werden können. Der Exponential-Mechanismus[9] hingegen kann genutzt werden, um synthetische Tabellen nach dem Vorbild der Originaltabelle zu erzeugen.[10] Der Exponential-Mechanismus ist jedoch mit sehr hohem Aufwand verbunden.

Ein weiteres Verfahren für Differential Privacy sind randomisierte Antworten,[11] welche bereits lange in sozialwissenschaftlichen Studien eingesetzt werden.[12] 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, sodass mit statistischen Methoden Erkenntnisse aus den Daten abgeleitet werden können. 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.

IV. Anonymisierung von Texten

Bei Textdokumenten wird zwischen der Metadatenebene, der Inhaltsebene und der Schreibstilebene unterschieden. Auf all diesen Ebenen können Personenbezüge vorhanden sein. Bevor auf die Anonymisierung hinsichtlich jeder einzelnen Ebene eingegangen wird, werden zunächst diese Begriffe geklärt. 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.

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 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 Medium, das z.B. ein Dateisystem oder eine E-Mail sein kann.

Metadaten bergen die Gefahr, dass sie oft vom Ersteller nicht wahrgenommen werden, jedoch 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.

2. Anonymisierung auf der Inhaltsebene

Inhaltsdaten enthalten oftmals Entitäten, die die Identität des Autors oder die von Dritten referenzieren können. Diese lassen sich anders als Metadaten nicht mit einfachen Mitteln entfernen, 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.[13] Anschließend können diese Verweise mit verschiedenen Strategien anonymisiert werden. Eine Garantie zur 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 Schlüssel k verschlüsselt, sodass aus dem Dokument D ein modifiziertes Dokument D' entsteht. 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 D' durch die verschlüsselten Elemente gestört wird.

Eine Alternative zur partiellen Verschlüsselung ist, die Verweise auf Identitäten zu paraphrasieren. Damit kann eine Anonymisierung erreicht und gleichzeitig die Semantik von D bis zu einem gewissen Grad beibehalten werden. Dazu gilt es, die identifizierten Entitäten durch generischere Angaben 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. Diese müssen in der Regel händisch erstellt werden. Alternativ eignen sich Ansätze basierend auf sog. 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 hinsichtlich einer Entität x semantisch ähnliche Entitäten y_1, y_2, ... finden, die eine Ersetzung erlauben. Vorausgesetzt werden hier jedoch genügend ungelabelte Textdaten, welche Informationen über die Entität x enthalten. Nachteil hierbei ist, dass die Entitäten nicht in einer festgelegten Relation (z.B. Synonymie) zueinander stehen, sondern sich über mehrere Relationen wie etwa Hyperonymie, Hyponymie oder Holonymie erstrecken können. Der Literatur zufolge existiert noch kein zufriedenstellender Ansatz, mit dessen Hilfe Entitäten hinsichtlich ihrer semantischen Relationen automatisiert abgegrenzt werden können.

3. Anonymisierung auf der Schreibstilebene

Die Identität einer Person lässt sich auch über deren 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.[14]

Aus der Notwendigkeit heraus, die Identität von Autoren zu schützen, entstand das Forschungsfeld Author Obfuscation (AO), welches sich mit der Verschleierung des Schreibstils in Dokumenten befasst. Bisherige AO-Ansätze lassen sich in manuelle, computerassistierte und automatische Verfahren aufteilen.[15] Automatische AO gilt als sehr anspruchsvoll, da sie auf Sprachkompetenzen zurückgreifen muss, um anonymisierende Umfor-

8 *Dwork* et al., in: Theory of Cryptography, 2006, p. 265.
9 *McSherry/Talwar*, in: FOCS 2007, p. 94.
10 *Blum/Ligett/Roth*, in: STOC 2008, p. 609.
11 *Kasiviswanathan* et al., in: FOCS 2008, p. 531.
12 *Warner*, JASA 60/309 (1965), 63.
13 *Li* et al., arXiv:1812.09449v1, 2018; *Yadav/Bethard*, in: COLING 2018, p. 2145.
14 *Potthast* et al., in: Advances in Information Retrieval, 2019, p. 291.
15 *Gröndahl/Asokan*, arXiv:1902.08939v2, 2019.

mungen 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^{4}NT)[16] hervorzuheben. Das Verfahren ist, soweit ersichtlich, das einzige Verfahren, das eine dedizierte Komponente für die Semantikerhaltung enthält. A^{4}NT 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^{4}NT das Dokument in dieselbe Sprache wie die Quellsprache „übersetzt", um den Schreibstil des ursprünglichen Autors nicht mehr wiedererkennen zu können. Das Verfahren wurde 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. Die Erkennungsgenauigkeit (F_1-Wert) sank beim Alter von 88% auf 8%, beim Geschlecht von 75% auf 39% und bei der Identität von 100% auf 0%, was dafür spricht, dass eine Anonymisierung auf der Schreibstilebene möglich ist.

V. Anonymitätsrisiken beim maschinellen Lernen

Da Algorithmen des maschinellen Lernens (ML) 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".[17] *Fredrikson* et al.[18] demonstrierten, dass die Erinnerung in neuronalen Netzen mitunter so stark sein kann, dass ein Abbild der Trainingsdaten rekonstruiert werden kann – ein sog. Modellinversionsangriff. *Shokri* et al.[19] bewiesen, dass neuronale Netze auf Grund 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. Daher kann ein Angreifer eindeutig zuordnen, ob ein Individuum in einem bestimmten Datensatz enthalten ist oder nicht. Solche Angriffe sind besonders dann kritisch, wenn es sich um sensible Informationen handelt, z.B. ob eine bestimmte Krankheit vorliegt.

Das Forschungsfeld Privacy Preserving Machine Learning (PPML) ist noch recht jung. Nachfolgend werden die wichtigsten Forschungsrichtungen auf diesem Gebiet skizziert.

1. Kollaboratives maschinelles Lernen

Das Ziel beim kollaborativen maschinellen Lernen ist, Daten auf privatsphärenfreundliche Weise einem ML-System zur Verfügung stellen zu können.

a) Kryptografische Verfahren für kollaboratives Lernen

Ein Ansatz, um die Privatheit des Einzelnen zu schützen und gleichzeitig das Training von Modellen auf Daten von mehreren Personen zu ermöglichen, ist die sichere Mehrparteienberechnung (engl. „secure multi-party computation" – MPC), ein Teilgebiet der Kryptografie. Das Ziel von MPC ist das gemeinschaftliche Berechnen einer Funktion, für die mehrere Parteien eine Eingabe liefern. Privatheit wird hierbei dadurch gewahrt, dass jede der beteiligten Parteien nur das Endergebnis, d.h. die Funktionsausgabe, erfährt, während die Eingaben der übrigen Teilnehmer verborgen bleiben. Tatsächlich gibt es erste MPC-Ansätze für gewisse Berechnungen im Kontext von maschinellem Lernen.[20]

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. Homomorphe Verschlüsselung ermöglicht daher, dass Daten zu Analysezwecken an eine nicht-vertrauenswürdige Instanz weitergegeben werden. Diese führt die Berechnungen auf den verschlüsselten Daten aus und übermittelt das verschlüsselte Ergebnis zurück. Insbesondere die sog. voll-homomorphe Verschlüsselung generiert jedoch einen signifikanten Rechenmehraufwand,[21] der diese für rechenintensive Anwendungen wie maschinelles Lernen bisher unbrauchbar macht. Erste praktikable Ansätze im Kontext von maschinellem Lernen verwenden daher Vereinfachungen.[22]

b) Dezentrales maschinelles Lernen

Beim dezentralen Lernen trainieren die Nutzer ein Grundmodell lokal auf ihren individuellen Daten und übermitteln lediglich die neu berechneten Parameter an den Service-Provider. 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 Parameter zurück an den Server.[23] Der Privatsphärenschutz kann gesteigert werden, wenn nicht alle Aktualisierungen mit dem Server geteilt werden.[24] Allerdings sollte sich der Nutzer des Trade-off zwischen der Menge der geteilten Aktualisierungen sowie der Trainingszeit und -qualität bewusst sein.

Hitaj et al.[25] haben nachgewiesen, dass es selbst in solchen dezentralen Lernansätzen mithilfe eines Generative Adversarial Network (GAN) möglich ist, über die übrigen aufrichtigen Teilnehmer sensible Daten zu sammeln. *Melis* et al.[26] entkräften teilweise die Angriffe von *Hitaj* et al., zeigen aber selbst neue Angriffsstrategien auf.

2. Differential Privacy für maschinelles Lernen

Arbeiten zu Differential Privacy (vgl. unter III.) im Kontext des maschinellen Lernens erforschen verschiedene Aspekte des Verrauschens von potenziell angreifbaren Daten. Untersucht wird meist, an welcher Stelle die Störungen Eingang in den Algorithmus finden sollten und welche Verteilungseigenschaften das Rauschen selbst haben sollte. 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 z.B. mittels Copula-Funktionen[27] oder Generative Adversarial Networks[28] 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. Darüber hinaus besitzen die Daten annähernd die gleichen Verteilungseigenschaften wie die zu Grunde liegenden Originaldaten und können in beliebiger Anzahl generiert werden, um so z.B. die Güte eines ML-Models zu verbessern.[29] Allerdings sto-

16 *Shetty/Schiele/Fritz*, in: USENIX Security 2018, p. 1633.
17 *Al-Rubaie/Chang*, IEEE Security & Privacy 17/2 (2019), 49.
18 *Fredrikson/Jha/Ristenpart*, in: CCS 2015, p. 1322.
19 *Shokri* et al., in: IEEE S&P 2017, p. 3.
20 *Bonawitz* et al., in: CCS 2017, p. 1175.
21 *Dowlin* et al., in: ICML 2017, p. 201; *Liu* et al., IEEE Access 6 (2018), 12103.
22 *Dowlin* (o. Fußn. 21); *Long* et al., arXiv:1811.10296v1, 2018.
23 *McMahan* et al., in: AISTATS 2017, p. 1273.
24 *Shokri/Shmatikov*, in: ACM CCS 2015, p. 1310.
25 *Hitaj/Ateniese/Pérez-Cruz*, in: CCS 2017, p. 603.
26 *Melis* et al., in: IEEE S&P 2019, p. 497.
27 *Li/Xiong/Jiang*, in: EDBT 2014, p. 475.
28 *Triastcyn/Faltings*, in: PAL, 2019, p. 33.
29 *McKay Bowen/Liu*, arXiv:1602.01063v4, 2019; *Page/Cabot/Nissim*, Differential privacy: an introduction for statistical agencies, 2018.

ßen die existierenden DIPS-Verfahren insbesondere bei hochdimensionalen Daten an Effizienzgrenzen.

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 auf Grund von höherem Berechnungsaufwand, extrem langen Trainingszeiten oder weil der Nutzen der Daten z.B. durch zugefügtes Rauschen vermindert wird. In manchen Fällen fallen diese Kosten sogar so groß aus, dass eine Anwendung in der Praxis nicht tragbar ist.[30]

VI. 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.

Für strukturierte Daten sind die heutigen Kernmethoden zur Anonymisierung hauptsächlich vor zehn bis zwanzig Jahren publiziert worden und seitdem vielfach weiterentwickelt worden. Handlungsbedarf besteht aber weiterhin in Bezug auf die Minimierung des Informationsverlusts bei gleichzeitiger Maximierung der Effizienz von Algorithmen.

Bei der Anonymisierung von Texten auf der Inhaltsebene stellt die zuverlässige Erkennung und Ersetzung von Entitäten nach wie vor eine Herausforderung dar. Bei der Anonymisierung der Stilebene gibt es erste Ansätze und empirische Ergebnisse, aber eine allgemeine Zuverlässigkeit kann noch nicht daraus geschlossen werden.

Der Privatsphärenschutz in Verbindung mit maschinellem Lernen ist ein junges Forschungsfeld. Erste Lösungsansätze existieren – etwa in Verbindung mit Differential Privacy oder Kryptografie. Neben den allgemeinen Herausforderungen dieser Schutzstrategien sind hier auch die Herausforderungen der Anwendbarkeit und Effektivität im Kontext des maschinellen Lernens zu lösen.

Abschließend lässt sich sagen, dass bei strukturierten Daten unter Berücksichtigung der Einschränkungen in Bezug auf Anonymitätsgarantien und Algorithmeneffizienz ein Praxiseinsatz von Anonymisierungsverfahren bereits möglich ist. In den anderen betrachteten Gebieten hingegen sind die vorgestellten Anonymisierungsstrategien hauptsächlich prototypische Forschungsarbeiten. In allen Bereichen gibt es noch viele zu lösende Forschungsfragen, von denen hier einige herausgearbeitet wurden.

Christian Winter
ist wissenschaftlicher Mitarbeiter am Fraunhofer-Institut für Sichere Informationstechnologie SIT in Darmstadt in der Abteilung Media Security and IT Forensics.

Verena Battis
ist wissenschaftliche Mitarbeiterin am Fraunhofer-Institut für Sichere Informationstechnologie SIT in Darmstadt in der Abteilung Media Security and IT Forensics.

Oren Halvani
ist wissenschaftlicher Mitarbeiter am Fraunhofer-Institut für Sichere Informationstechnologie SIT in Darmstadt in der Abteilung Media Security and IT Forensics.

Dieser Beitrag enthält Ergebnisse aus dem Teilprojekt „Privacy und Big Data“ i.R.d. vom Hessischen Ministerium des Innern und für Sport geförderten Projekts „Cybersicherheit für die digitale Verwaltung“.

30 *Al-Rubaie/Chang* (o. Fußn. 17).