Fakultät für Mathematik und Informatik
Ruprecht-Karls-Universität Heidelberg

Differential Privacy

Seminar „Ist künstliche Intelligenz gefährlich?“

Sommersemester 2017

Ausarbeitung von Mira Boehme
E-Mail: mira.boehme@web.de
Matrikelnummer.: 3055868

Studiengang: Informatik Lehramt
Leitung: Prof. Dr. Ullrich Köthe
Abgabedatum: 09. September 2017

Inhaltsverzeichnis

1 Einführung

Datenschutz ist in der heutigen Informationsgesellschaft ein äußert wichtiges Thema. Große Datenkraken wie Google oder Facebook nutzen jede Gelegenheit, um personenbezogene Daten zu erheben und dadurch möglichst genaue Profile ihrer Nutzer zu erstellen. Die eigene Privatsphäre im Internet zu schützen ist keine einfache Angelegenheit, denn Schutzsysteme wie Firewalls sind hauptsächlich ein Schutz gegen Schadsoftware wie Viren und Trojaner. Als Nutzer von Internetdiensten muss man sozusagen darauf hoffen, dass die eigenen Daten vertraulich behandelt und im besten Fall nur verschlüsselt oder anonymisiert weitergegeben und verarbeitet werden. Konzepte dazu gibt es sehr wohl – beispielweise die „homomorphe Verschlüsselung“ oder das Konzept der „Differential Privacy“. Letzteres soll in dieser Arbeit vorgestellt und erklärt werden.

Zunächst wird in Kapitel 1 ein kleiner Überblick über verschiedene Möglichkeiten, Datenschutz zu gewährleisten, gegeben und das Prinzip der Differential Privacy wird kurz erläutert. In Kapitel 2 sollen zwei Beispiele, bei denen versuchter Datenschutz schiefgegangen ist, als Motivation für die Differential Privacy gegeben werden. Kapitel 3 geht genauer auf das Konzept der Differential Privacy ein; hier werden die Grundidee und das Vorgehen erläutert. In Kapitel 4 wird zunächst die Entwicklerin des Konzepts – Cynthia Dwork – kurz vorgestellt und danach werden die formalen Definitionen der Differential Privacy und Beispiele dazu gegeben. Auf verschiedene Mechanismen, mit denen Differential Privacy umgesetzt werden kann, wird in Kapitel 5 eingegangen. In Kapitel 6 werden die Stärken und Schwächen des Konzepts erläutert und in Kapitel 7 erfolgt ein kurzes Fazit.

1.1 Apple und der Datenschutz

Auf der „Worldwide Developers Conference“ im Juni 2016 kündigte Apple an, in Zukunft auf das Konzept der Differential Privacy zurückgreifen zu wollen. Mit dem neuen Betriebssystem iOS 10 sollen mehr Nutzerdaten gesammelt werden, um Dienste besser auf einzelne Nutzer zuschneiden zu können. Allerdings soll dabei trotzdem die Privatsphäre jeder beteiligten Person geschützt werden (vgl. [2] S. 41).

Dieser Kompromiss zwischen einerseits Daten zu sammeln und zu verarbeiten, aber andererseits auch Daten zu schützen, ist mittlerweile von sehr großer Bedeutung. Unternehmen brauchen tatsächlich viele Daten, um ihre Dienste laufend zu verbessern, sollten aber stets darauf bedacht sein, dabei die Privatsphäre ihrer Nutzer zu wahren. Ein Konzept, welches diesen Kompromiss umsetzt, ist die Differential Privacy.

1.2 Verschlüsselung von Daten

Auch die Verschlüsselung von Daten ist eine Möglichkeit, die Privatsphäre von Nutzern zu schützen. Allerdings ist Verschlüsselung nicht immer zielführend, denn die Verarbeitbarkeit der Daten wird hierbei eingeschränkt. Mit Hilfe von homomorpher Verschlüsselung können zwar Operationen auf verschlüsselten Daten ausgeführt werden, generell ist dies aber relativ teuer. Bei den meisten anderen Verschlüsselungsverfahren müssen Daten erst wieder entschlüsselt werden, bevor Operationen auf ihnen durchgeführt werden können.

Verschlüsselung verhindert außerdem die uneingeschränkte Veröffentlichung von Daten, was für gewisse Zwecke allerdings notwendig ist. Beispielsweise wäre es wenig zielführend, wenn ein Krankenhaus seine Patientendaten zu Forschungszwecken verschlüsselt weitergibt, denn Forscher müssen auf gewisse Daten uneingeschränkt zugreifen können, um einen Nutzen aus ihnen ziehen zu können. Deshalb wird ein Verfahren benötigt, um Daten veröffentlichen zu können, ohne dabei den Datenschutz zu verletzen. Dies kann durch das Anonymisieren von Daten erreicht werden (vgl. [2] S. 28).

1.3 Anonymisieren von Daten

Bei der Anonymisierung von Daten unterscheidet man drei verschiedene Arten von Daten: die (direkten) Identifikatoren, die Quasi-Identifikatoren und die sensiblen Werte. Direkte Identifikatoren sind Angaben wie der volle Name einer Person oder deren Personalausweisnummer, mit welchen eine Person direkt identifiziert werden kann. Als Quasi-Identifikatoren bezeichnet man Kombinationen aus Attributen wie beispielsweise das Tripel „Postleitzahl, Geburtsdatum, Geschlecht“, welches einer Person eindeutig zugeordnet werden kann. In kleineren Städten wie Heidelberg kann dieses Attribut-Tripel durchaus dazu dienen, Personen zweifelsfrei zu identifizieren. Die sensiblen Werte sind

Attribute, die man aus unterschiedlichen Gründen nicht mit einzelnen Personen in Verbindung bringen können sollte (vgl. [2] S. 29).

Für das Anonymisieren von Daten gilt es gewisse Regeln und Richtlinien zu beachten, um zu gewährleisten, dass nicht trotz Anonymisierung Rückschlüsse auf Personen gezogen werden können. Allerdings können „Angreifer“, die ein Interesse an sensiblen Daten haben, gewisses Hintergrundwissen besitzen, mithilfe dessen trotz aller Vorsicht die De-Anonymisierung von Daten gelingen kann. Die Bandbreite an möglichem Hintergrundwissen ist enorm groß und dieses Risiko kann daher kaum realistisch eingeschätzt, geschweige denn eliminiert werden. Anonymisieren reicht also nicht unbedingt aus, um Datenschutz zu gewährleisten. Das Konzept der Differential Privacy hingegen kann Datenschutz beliebig sicher gewährleisten, wie sich in Kapitel 4 zeigen wird (vgl. [2] S. 38).

2 Motivation

Im Folgenden werden als Motivation zwei Beispiele erläutert, in denen das Anonymisieren von Daten nicht den gewünschten Zweck erfüllte. Das erste Beispiel behandelt den sogenannten „Netflix Prize“, der vor ungefähr zehn Jahren für Aufsehen sorgte. Im zweiten Beispiel wird ein Vorfall aus Massachusetts erläutert, der aufgrund von Unachtsamkeit der „Massachusetts Group Insurance Commission“ einen kleinen Skandal auslöste.

2.1 „Netflix Prize“

Der „Netflix Prize“ war ein im Jahr 2006 vom Streaming-Dienst Netflix ausgeschriebener Wettbewerb. Er versprach eine Millionen US Dollar für denjenigen Teilnehmer, der den besten „Vorhersage-Algorithmus“ einreichen würde. Hintergrund des Wettbewerbs war, dass Netflix die Film-Empfehlungen, die seinen Nutzern gemacht wurden, verbessern wollte. Wenn nämlich das Verhalten einzelner Nutzer (hier in Bezug auf Bewertungen) präzise vorhergesagt werden kann, können bessere Empfehlungen gemacht werden (vgl. [2] S. 29 und [3] S. 2).

Netflix stellte über 500.000 Datensätze von Nutzern bereit, mit deren Hilfe die Programmierer ihre Algorithmen trainieren konnten. Diese Trainings-Datensätze hatten die Form < user, movie, date of grade, grade >. Ein Datensatz bestand also aus einem Nutzernamen, einem Filmtitel, der Bewertung des Nutzers für den jeweiligen Film und dem Datum der Bewertung. Die Datensätze wurden natürlich anonymisiert herausgegeben, indem für jeden Nutzernamen ein zufälliges Pseudonym erstellt wurde (vgl. [2] S. 29 und [3] S. 2).

Neben dem Filmbewertungssystem von Netflix gibt es noch weitere Onlinedienste, in denen Filme bewertet werden können wie beispielsweise die „Internet Movie Database“ (IMDb). Bei IMDb gibt es einige Nutzer, die mit ihrem echten Namen angemeldet sind. Zwei Wissenschaftler der „University of Texas“ kamen daher auf die Idee, die Trainings-Datensätze von Netflix mit den Bewertungen aus der IMDb abzugleichen. Dadurch konnten tatsächlich einige Nutzer aus den Netflix-Datensätzen identifiziert werden und es gelang eine partielle De-Anonymisierung der Daten (vgl. [2] S. 30 und [3] S. 2).

2.2 „Massachusetts-Missgeschick“

Die “Massachusetts Group Insurance Commission” (GIC) ist die Krankenkasse von über 135.000 Staatsbediensteten und deren Angehörigen in Massachusetts. Zu Forschungszwecken gab die GIC – natürlich anonymisiert – Patientendaten an die Industrie und an Forscher weiter. Diese beinhalteten neben Gesundheitsdaten wie Diagnosen und Arztbesuche auch persönliche Daten wie Postleitzahl, Geburtsdatum und Geschlecht der Patienten (vgl. [2] S. 31).

Für Cambridge in Massachusetts gibt es ein öffentliches Wählerverzeichnis, welches für 20 $ zu erwerben ist und welches neben Parteizugehörigkeiten auch persönliche Daten wie das Geburtsdatum eines Wählers offenlegt. Die Professorin Sweeney der „Pittsburgh-University“ begann, die Patientendaten der GIC mit diesem Wählerverzeichnis abzugleichen und ihr gelang es dadurch einige Personen zu reidentifizieren. Durch den Quasi-Identifikator < PLZ, Geburtsdatum, Geschlecht > (Abbildung 2.1) konnte sie unter anderen William Held, den ehemaligen Gouverneur von Massachusetts, identifizieren, denn es gab nur eine männliche Person mit seinem Geburtsdatum und seiner Postleitzahl. Professor Sweeney bekam durch ihre Nachforschungen also sozusagen Einsicht in seine

Krankenakte, was damals für viel Aufsehen sorgte und die GIC in ein schlechtes Licht rückte (vgl. [2] S. 31).

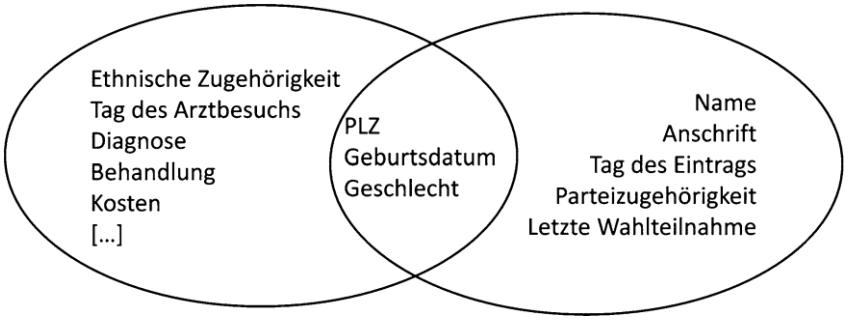

Abbildung 2.1: Quasi-Identifikator ([2] S. 31)

Eine Studie aus dem Jahre 2006 kam zu der Erkenntnis, dass sogar 63,3 % der amerikanischen Bevölkerung durch diese Attributkombination (PLZ, Geburtsdatum, Geschlecht) eindeutig identifizierbar seinen (vgl. 14[2] S. 31).

3 Konzept der Differential Privacy

"Imagine you have two otherwise identical databases, one with your information in it, and one without it. Differential Privacy ensures that the probability that a statistical query will produce a given result is (nearly) the same whether it's conducted on the first or second database." (Green, 2016 in [5])

Dieses Zitat des Informatikers Matthew Green beschreibt das Grundprinzip der Differential Privacy recht gut. Es wird eine statistische Garantie darüber gegeben, dass die Daten einzelner Personen keine Auswirkung auf das Ergebnis bestimmter Abfragen haben. Im folgenden Abschnitt wird dieses Grundprinzip näher erläutert und in Abschnitt 3.2 das Vorgehen beschrieben, wie Differential Privacy umgesetzt werden kann.

3.1 Grundidee

Die Differential Privacy gibt ein Maß dafür, wie hoch das Risiko für eine einzelne Person ist, an einer statistischen Datenbank teilzunehmen. Als Grundsatz gilt, dass personenbezogene Daten zu keinem Unterschied im Ergebnis einer Studie führen dürfen. Denn die Wahrung der Privatsphäre einer Person ist genau dann gewährleistet, wenn das

Ergebnis einer Datenbank-Abfrage nicht von den Daten einer einzelnen Person abhängt (vgl. [2] S. 38).

3.2 Vorgehen

Um Differential Privacy gewährleisten zu können, werden Funktionen benötigt, die Datenbankabfragen beantworten können und dabei sicherstellen, dass der Datenschutz nicht verletzt wird (vgl. [2] S. 38). Dies geschieht, indem die Originaldaten während einer Abfrage mit sogenanntem „Rauschen“ versehen werden. Je nach Abfrage werden die Daten zufällig abgeändert, indem beispielsweise neue Einträge – sogenannte „Dummys“ – generiert oder gewisse Attribute vertauscht werden. Wichtig ist aber, dass die abgeänderten Daten nicht von den ursprünglichen Daten zu unterscheiden sind, denn sonst können diese als Originaldaten erkannt werden und dadurch wieder Rückschlüsse auf Personen gezogen werden. Bei der Generierung neuer Einträge dürfen außerdem keine statistischen Zusammenhänge verfälscht werden. Abgeänderte Daten, die Auswirkungen auf das Ergebnis einer Datenbankabfrage haben, sind zu Forschungszwecken unbrauchbar (vgl. [7] und [3] S. 2).

4 Formale Definition

Zu beachten ist, dass Differential Privacy kein bestimmter Algorithmus ist, sondern eine Definition. Wie diese genau aussieht, wird in den Abschnitten 4.2 und 4.3 dargestellt. Zunächst soll aber kurz die Entwicklerin der Differential Privacy vorgestellt werden.

4.1 Cynthia Dwork

Cynthia Dwork gilt als die Entwicklerin der Differential Privacy. Geboren wurde sie 1958 in den USA als Tochter eines bekannten Mathematikers. Im Jahre 1983 erlangte sie an der Cornell University ihren Doktortitel in Informatik im Bereich „Distributed Computing“. Dwork forscht unter anderem in Kryptographie und seit 2017 ist sie Gordon-McKay-Professorin für Informatik an der „Harvard University“. Für ihre

Forschungsarbeit erhielt sie im Jahre 2007 den Dijkstra-Preis. Das Konzept der Differential Privacy entwickelte sie gemeinsam mit drei anderen Informatikern. Im Jahre 2006 wurde es der Öffentlichkeit präsentiert und mittlerweile findet es auch Anwendung bei namhaften Firmen wie Apple (vgl. [4]).

4.2 ε-Differential Privacy

Die formale Definition der ε-Differential Privacy wirkt auf den ersten Blick sehr technisch. Sie besagt, dass eine randomisierte Funktion κ ε-DP bietet, wenn

- für alle Datensätze D_1 und D_2, die sich nur in höchstens einem Element unterscheiden
- für alle Teilmengen S des Wertebereichs W von κ gilt:

$$P[\kappa(D_1) \in S] \leq e^{\varepsilon} * P[\kappa(D_2) \in S]$$

Die Wahrscheinlichkeit, dass eine Datenbankabfrage über einer Datenbank D_1 ein gewisses Ergebnis liefert, soll multiplikativ gleich sein wie die Wahrscheinlichkeit derselben Abfrage über der Datenbank D_2. Hierbei dürfen sich die beiden Datensätze in nur einem Element unterscheiden, einer der beiden Datensätze darf also höchstens einen Eintrag mehr enthalten. Die anderen Elemente müssen identisch sein (vgl. [2] S. 38).

Als Beispiel betrachten wir Abbildung 4.1, die eine Tabelle mit Patientendaten eines Krankenhauses enthält. In dieser Auflistung sind vier männliche Patienten und drei weibliche Patienten enthalten. Fügen wir nun dieser Tabelle einen weiteren Datensatz eines männlichen Patienten hinzu, unterscheidet sich die neue Tabelle in genau einem Element von der ursprünglichen Tabelle. Eine einfache Zählabfrage, die die Anzahl der männlichen Patienten ausgeben soll, würde nun über die Originaltabelle die Antwort „4“ erzeugen, über die neue Tabelle allerdings lautet die Antwort „5“. Diese beiden Antworten sind zu 100% sicher, eine andere Antwort ist nicht möglich. Die Funktion liefert also über die beiden Datensätze in jedem Fall ein unterschiedliches Ergebnis und bietet somit keine ε-DP (vgl. [2] S. 39).

Geburtsjahr	PLZ	Geschlecht	Diagnose
1982	33098	Männlich	Migräne
1982	33098	Männlich	Erkältung
1983	33098	Männlich	Rheuma
1983	33098	Männlich	Depression
1985	33100	Weiblich	Heuschnupfen
1985	33100	Weiblich	Hypochondrie
1983	33098	Weiblich	Migräne

Abbildung 4.1: Patientendaten

Um in diesem Fall ε-DP zu gewährleisten, wäre eine mögliche Methode, die Zählabfrage so abzuändern, dass sie die männlichen Patienten nur zu einer gewissen Wahrscheinlichkeit als Mann erkennt. Dadurch kann die multiplikative Gleichheit der Wahrscheinlichkeiten über die beiden Datensätze erreicht werden. Diese Methode, ε-DP zu erreichen, nennt man „Randomized Response"-Mechanismus (vgl. [2] S. 39). Hierauf wird in Kapitel 5.3 näher eingegangen.

Der Faktor ε gibt an, wie groß die Garantie auf Datenschutz ist. Je kleiner der Faktor ist, desto höher die Garantie, die gegeben werden kann. Dadurch wird allerdings generell der Nutzen der Daten geringer. In der Praxis gibt es sowohl sehr große als auch sehr kleine ε. Dies hängt von der Wichtigkeit und der Sensibilität der Daten ab (vgl. [2] S. 39).

Die ε-DP birgt außerdem gewisse Einschränkungen, was die Anwendung betrifft. So können Ergebnisse, die mit Wahrscheinlichkeit 0 eintreffen, nicht multiplikativ beeinflusst werden. Daher gibt es eine Weiterentwicklung der ε-DP um einen additiven Parameter δ, die (ε, δ)-Differential Privacy (vgl. [7]).

4.3 (ε, δ)-Differential Privacy

Die Erweiterung der (ε, δ)-Differential Privacy um den additiven Parameter δ erlaubt, dass Voraussetzungen bis zu einem gewissen Grad δ unerfüllt bleiben. δ sollte möglichst klein gewählt werden, denn sonst sind die veränderten Daten möglicherweise sehr einfach auf die Originaldaten zurückzuführen. Wie klein δ gewählt werden sollte, um Datenschutz noch garantieren zu können, hängt von der Größe der Datenbank ab. Die Definition der (ε, δ)-Differential Privacy lautet:

Eine randomisierte Funktion κ bietet (ε, δ)-Differential Privacy, wenn

- für alle Datensätze D_1 und D_2, die sich nur in höchstens einem Element unterscheiden
- für alle Teilmengen S des Wertebereichs W von κ gilt:

$$P[\kappa(D_1) \in S] \leq e^{\varepsilon} * P[\kappa(D_2) \in S] + \delta$$

Sie unterscheidet sich also tatsächlich nur um den additiven Parameter δ von der Definition der ε-Differential Privacy (vgl. [7]).

5 Mechanismen

Im folgenden Kapitel werden zwei Mechanismen vorgestellt, die verwendet werden, um Differential Privacy zu erreichen. Einer ist der bereits angesprochene „Randomized Response"-Mechanismus, der ursprünglich für psychologische Zwecke entwickelt wurde. Der andere ist der Laplace-Mechanismus, bei welchen die Laplace-Verteilung eine wichtige Rolle spielt. Zunächst wird allerdings der Begriff der „Sensibilität" eingeführt, da dieser für einige der Mechanismen benötigt wird.

5.1 Sensibilität

Die Sensibilität beschreibt, wie sehr die Daten einer einzelnen Person das Ergebnis einer Datenbankabfrage beeinflussen können. Genau diese Information wollen wir mit Differential Privacy verbergen und daraus können wir wiederum ableiten, wie viel Rauschen zu den Daten hinzuzufügen ist. Die Sensibilität Δf einer Funktion f ist definiert durch:

$$\Delta f = \max_{D1,D2} |f(D_1) - f(D_2)| \text{ für „benachbarte" Datensätze } D_1, D_2$$

Die Sensibilität ist also der maximale Unterschied, den eine Datenbankabfrage über zwei verschiedene Datensätze erhalten kann. Die Datensätze sollen sich hierbei wieder nur in maximal einem Element unterscheiden. Die Sensibilität einer einfachen Zählabfrage beispielweise wäre 1, denn zwei benachbarte Datensätze können sich in einem bestimmten Attribut höchstens um einen Zähler unterscheiden (vgl. [7] und [1] S. 5).

5.2 Laplace-Mechanismus

Der Laplace-Mechanismus verfremdet die Originaldaten, indem ihnen kontrolliert Rauschen hinzugefügt wird. Dieses Rauschen hängt von der Sensibilität der jeweiligen Funktion ab und hat die mathematischen Eigenschaften einer Laplace-Verteilung (Abbildung 5.1). Deshalb wird es in diesem Fall Laplace-Rauschen Lap(b) genannt. Hierbei gilt $b = \Delta f/\varepsilon$. Bei einer Zählabfrage mit Sensibilität 1 wäre also $b = 1/\varepsilon$ und somit wäre den Daten Laplace-Rauschen der Form Lap(1/ε) hinzuzufügen (vgl. [7]).

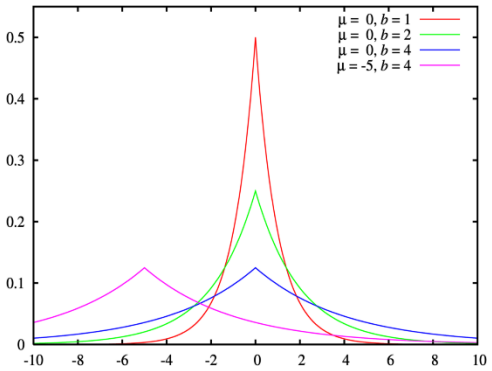

Abbildung 5.1: Schaubild der Laplace-Verteilung [6]

Generell muss beim Hinzufügen von Rauschen natürlich gewährleistet sein, dass die statistischen Verteilungen von möglichen Abfrageergebnissen nicht verfälscht werden. Die ursprünglichen Daten dürfen also nur so verändert werden, dass dies keine gravierenden Auswirkungen auf die Ergebnisse hat (vgl. [7]).

Beim Hintereinanderausführen von unterschiedlichen Mechanismen, die ε-DP sind, werden die jeweiligen Parameter ε_i verknüpft, um die resultierende ε-Differential Privacy zu erhalten. Führt man also zwei DP-Mechanismen hintereinander aus, wobei der eine ε_1-DP ist und der andere ε_2-DP, erhält man dadurch einen $(\varepsilon_1 + \varepsilon_2)$-DP-Mechanismus. Dies funktioniert auch bei (ε, δ)-Differential Privacy (vgl. [7]).

5.3 „Randomized Response“-Mechanismus

Das Konzept des „Randomized Response“-Mechanismus wird hauptsächlich in sozialwissenschaftlichen Studien verwendet. Hier dient es dazu, Rückschlüsse auf

einzelne Personen zu verhindern. Teilnehmer einer solchen Studie können mithilfe des Mechanismus brisante Fragen ehrlich beantworten, ohne Angst vor möglichen Konsequenzen zu haben. Brisante Fragen sind beispielsweise Fragestellungen wie „Hast du schon einmal etwas geklaut?“ oder „Bist du schon einmal betrunken Auto gefahren?“. Solche Fragen würden viele Personen nicht unbedingt ehrlich beantworten, da dies für sie ernsthafte Konsequenzen haben könnte. Der „Randomized Response“-Mechanismus wirkt dem entgegen, indem er verhindert, dass Antworten einer einzelnen Person zweifelsfrei zugeordnet werden können (vgl. [2] S. 39 und [3] S. 2).

Es gibt verschiedene Varianten des „Randomized Response“-Mechanismus. Die gängigste Variante basiert auf einem einfachen Münzwurf. Der Befragte wirft eine Münze. Ist das Ergebnis Zahl, beantwortet er die Frage wahrheitsgemäß. Ist das Ergebnis allerdings Kopf, wirft er die Münze noch einmal. Diesmal gilt: Bei Zahl antwortet der Befragte mit „Nein“, bei Kopf mit „Ja“ (vgl. [3] S. 2).

Es bleibt also ungewiss, ob der Befragte die Frage ehrlich beantwortet oder ob er die Antwort der Münze gibt. Trotz Zufälligkeit der Antworten verfälscht der Mechanismus nicht die statistischen Ergebnisse, denn die ehrlich gegebenen Antworten lassen sich aus dem Gesamtergebnis herausrechnen. Dies liegt daran, dass sich Zufallsexperimente wie der Münzwurf bei häufigen Wiederholungen stets dem erwarteten Ergebnis annähern und man somit davon ausgehen kann, dass die Hälfte der Befragten eine ehrliche Antwort gibt (vgl. [3] S. 2).

6 Stärken und Schwächen

Wie jedes Verfahren hat auch die Differential Privacy ihre Stärken und Schwächen. Als Stärke ist zu nennen, dass die DP ein tatsächliches Maß für Datenschutzgarantie geben kann, was beispielweise durch Verschlüsselungs- oder Anonymisierungsverfahren nicht unbedingt gegeben werden kann. Außerdem ist die DP-Methode nicht abhängig von der Computer-Power der Gegner, denn hier geht es nicht darum, wer den schnellsten und mächtigsten Algorithmus hat (vgl. [7]).

Allerdings stellt das Konzept der Differential Privacy sehr hohe Anforderungen an seine Mechanismen und ist daher recht aufwändig zu realisieren. Davon abgesehen können Daten auch schnell an Nutzen verlieren, wenn zu viel Rauschen generiert werden muss

und die Daten zu sehr von den Originaldaten abweichen. Außerdem können Anfragen an eine Datenbank nur so lange mit hoher Genauigkeit beantwortet werden, bis weitere Ausgaben den Datenschutz verletzen würden. Diese Menge an Informationen, die herausgegeben werden kann, wird als Privacy Budget bezeichnet (vgl. [3] S. 3).

7 Fazit

Für den Verbraucher wäre es sicherlich gut, wenn mehr Unternehmen Apples Beispiel folgen und Differential Privacy einsetzen würden. Aber es gibt natürlich nicht *das* Verfahren zum Datenschutz. Wichtig ist generell, dass Unternehmen zeigen, dass ihnen die Privatsphäre ihrer Nutzer am Herzen liegt, was längst nicht selbstverständlich ist. Gerade kleine Unternehmen haben nicht unbedingt die Mittel, Differential Privacy umzusetzen. Beispielsweise sind die Mechanismen, um Patientenakten mithilfe von Differential Privacy weitergeben zu können, doch sehr kompliziert und für ein Krankenhaus wären der Aufwand und die Kosten hierfür unverhältnismäßig groß.

8 Literaturverzeichnis

[1] Ji, Zhanglong, Lipton, Zachary C. und Elkan, Charles. *Differential Privacy and Machine Learning: a Survey and Review.* 2014.

[2] Petrlic, Ronald und Sorge, *Christoph. Datenschutz: Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie.* Springer Vieweg, 2017.

[3] https://www.heise.de/mac-and-i/artikel/Besserer-Datenschutz-Wie-Apples-Differential-Privacy-funktioniert-Update-3678489.html

[4] https://www.seas.harvard.edu/directory/dwork

[5] https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/

[6] http://de.academic.ru/dic.nsf/dewiki/828765

[7] https://www.youtube.com/watch?v=ekIL65D0R3o
Tutorial on Differential Privacy by Katrina Ligett, 2013.

9 Abbildungsverzeichnis