DAV/DGVFM
Jahrestagung
2022

Pseudonymisierung / Anonymisierung
Ein Überblick für Aktuar*innen

Dr. Felix Spangenberg und Dariush Sadeghi-Yam

DAV
DEUTSCHE AKTUARVEREINIGUNG e.V.

DGVFM
DEUTSCHE GESELLSCHAFT FÜR VERSICHERUNGS- UND FINANZMATHEMATIK e.V.

DAV/DGVFM-Jahrestagung, 27.–29. April 2022

Inhalt

Inhalt

Die AG Daten und Datenschutz

Mitglieder

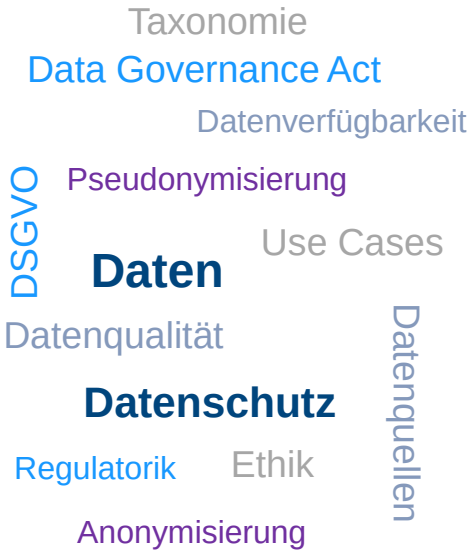
Christian Bökenheide
Thomas Lorentz
Bartek Maciaga
Tobias Renner
Dariush Sadeghi-Yam
Dr. Felix Spangenberg
Dr. Dirk Wehrmann
Barbara Winter

AG Daten und Datenschutz

„Ziel der Arbeitsgruppe ist es, sich mit der Erarbeitung von Grundlagen zur Datengewinnung und -aufbereitung sowie dem Datenschutz auseinanderzusetzen.

Neben Gesetzen und Richtlinien wird sich die Arbeitsgruppe auch mit Fragen der Ethik im Umgang mit Daten befassen."

- Der Aktuar 01/2019 -



… und die zugehörigen Pools

Anonymisierung / Pseudonymisierung

Aufbereitung von Grundlagen zur Anonymisierung und Pseudonymisierung

Dr. Stefan Karrmann
Dariush Sadeghi-Yam
Olga Schäfer[1]
Prof. Dr. Josef Schürle
Dr. Felix Spangenberg

- ☐ Ergänzung von Ausbildungsunterlagen
- ☐ Erstellung von Schulungsunterlagen
- ☐ Veröffentlichung in der Zeitschrift „Der Aktuar"

Analyse Use Cases im Hinblick auf den Datenschutz

Analyse von Use Cases aus verschiedenen Sparten der Versicherungswirtschaft hinsichtlich des Datenschutzes

Dariush Sadeghi-Yam
Olga Schäfer[1]
Dr. Christian Weiner

- ☐ Veröffentlichung in der Zeitschrift „Der Aktuar"

Mitglieder

[1]ausgetreten Ende 2021

Inhalt

Begriffsbestimmungen

Pseudonymisierung (Art. 4 DSGVO)

- Verarbeitung personenbezogener Daten
- ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zuordenbar
- zusätzliche Informationen werden gesondert aufbewahrt und unterliegen technischen und organisatorischen Maßnahmen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden

Anonymisierung

- Verarbeitung personenbezogener Daten
- Personenbezug kann nicht oder nur unter unverhältnismäßigem Aufwand wiederhergestellt werden
- „=" Pseudonymisierung mit Löschen der zusätzlichen Information zur Identifizierung

Besonders die Anonymisierung der Daten spielt im Kontext geplanter europäischer Verordnungen, wie dem Data Governance Act, Data Act oder Artificial Intelligence Act ein wichtige Rolle.

Arten von personenbezogenen Daten

Personenbezogene Daten für die Pseudonymisierung / Anonymisierung sind beispielsweise

- Name
- Geburtsdatum
- Ausweisnummer
- Persönliche Merkmale
- Biometrische Daten
- Patientendaten
- Finanzdaten
- Demographische Daten

Unterteilung in

(Explizite) Identifikatoren

Attribute, die eine eindeutige Identifikation von Personen ermöglichen

- Name
- Ausweisnummer

Quasi-Identifikatoren

Menge von (nicht sensitiven) Attributen, die durch die Hinzunahme externer Daten es ermöglichen, eine Person zu identifizieren

- Kombination aus Geschlecht, Postleitzahl und Alter

Sensitive Attribute

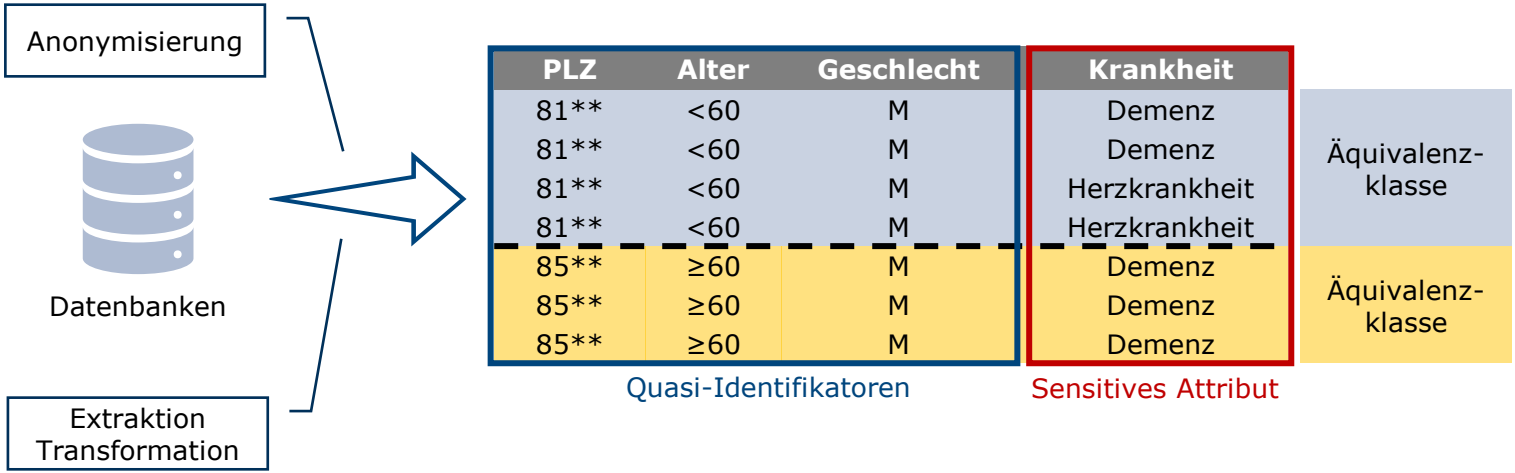
Persönliche und schützenswerte Informationen

- Gehalt
- Krankheit

Äquivalenzklassen

Zusammenfassung von Daten zu **Äquivalenzklassen**:

Die Daten innerhalb eines Datensatzes, deren Quasi-Identifikatoren dieselben Werte annehmen, werden in einer Äquivalenzklasse zusammengefasst.



Inhalt

„Einfache“ Anonymisierungsmethoden

Generalization

Eine **Generalisierung** verändert Werte von Quasi-Identifikatoren, um ihre Genauigkeit zu reduzieren. Kategoriale Werte können anhand einer Taxonomie durch allgemeinere Werte ersetzt werden. Bei numerischen Attributen werden exakte Angaben durch Intervalle ersetzt.

Suppression

Suppression unterdrückt bestimmte Werte bzw. ersetzt sie durch ein spezielles Zeichen, sodass sie im resultierenden Datensatz nicht mehr vorhanden sind.

Perturbation

Bei der **Perturbation** werden die Datenwerte durch künstlich generierte Werte ersetzt. Das Ziel dabei ist, die Daten so zu verändern, dass dennoch statistische Eigenschaften des Datensatzes für Analysen erhalten bleiben.

Permutation

Bei der **Permutation** werden Daten zwischen den Datensätzen innerhalb einzelner Attribute gemischt. Grundlage sollte eine Zufallsverteilung sein, dabei ist grundsätzlich nicht auszuschließen, dass ein Datensatz auf sich selbst abgebildet wird (ggf. durch Vorkehrungen auszuschließen).

Generalization Beispiel

Postleitzahl

- 81*
  - 811*
  - 812*
  - 813*

Beruf

- Künstler*in
  - Sänger*in
  - Maler*in

Krankheit

- Atemwegs-erkrankung
  - Asthma
  - Bronchitis

Suppression Beispiel

Value Suppression

Name	Bundesland	Herkunft	Alter	Geschlecht	Krankheit
*	BY	*	65	M	Demenz
*	HE	*	72	F	Herzkrankheit
*	*	*	*	*	*
*	NI	*	45	M	Herzkrankheit
*	NW	*	32	*	Krebs
*	HE	*	55	F	Herzinfarkt
*	*	*	61	D	Bronchitis
*	SA	*	59	M	Alzheimer

Record Suppression

Cell Suppression

Permutation - Beispiel

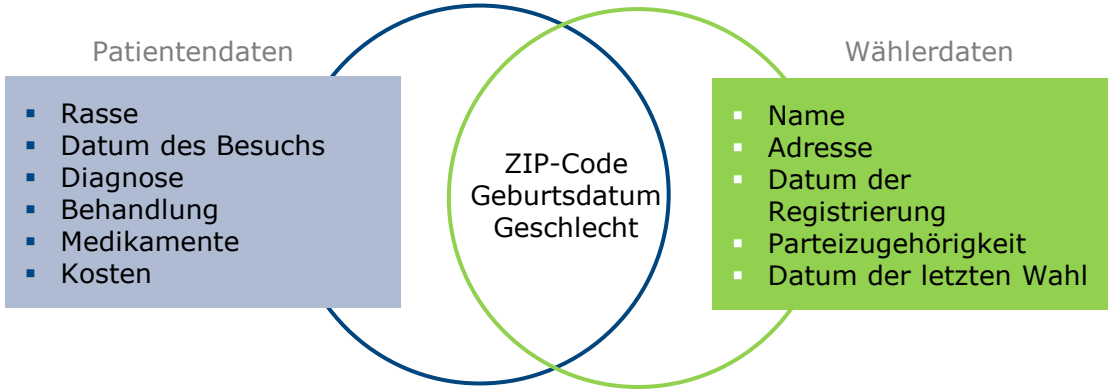
Alter	Geschlecht	Krankheit
65	M	Demenz
72	F	Herzkrankheit
53	M	Schizophrenie
45	M	Herzkrankheit
32	F	Krebs
55	F	Herzinfarkt
61	D	Bronchitis
59	M	Alzheimer

Alter	Geschlecht	Krankheit
65	M	Herzkrankheit
72	F	Alzheimer
53	M	Krebs
45	M	Schizophrenie
32	F	Bronchitis
55	F	Herzkrankheit
61	D	Demenz
59	M	Herzinfarkt

Inhalt

Wie anonym sind Daten? – Ein Beispiel

- In Massachusetts ist die Group Insurance Commission (GIC) dafür verantwortlich, Versicherungen für die Staatsangestellten zu erwerben.
- Dabei sammelt GIC verschiedene personenbezogene Daten (z.B. Patientendaten). Diese Daten wurden nach Anonymisierung der Forschung und der Industrie zur Verfügung gestellt.
- Durch Hinzubinden der Informationen aus Wählerregistrierlisten zeigt L. Sweeney[1] am Beispiel des Gouverneurs von Massachusetts, wie personenbezogene Daten offengelegt werden können.



[1]Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*, 05 (2002), 557–570.

Metriken für Anonymität von Datensätzen

- Methoden zur Anonymisierung haben das Ziel, den Personenbezug aus Daten zu entfernen
- Es stellt sich jedoch die Frage, „wie anonym" die Daten sind.

Sind einzelne Personen mit großer Wahrscheinlichkeit identifizierbar?

Unabhängig von den sensiblen Informationen, wie viele Personen mit den gleichen Merkmalen gibt es?

Auch wenn Personen mit den gleichen Merkmalen existieren, in wie weit unterscheiden sich die sensiblen Informationen?

⇨ Metriken zur Messung von Anonymität von Datensätzen sind u.a.
- k-Anonymität (siehe nächste Folie)
- l-Diversität
- t-Ähnlichkeit

k-Anonymität

- Definition: Ein Datensatz genügt genau dann der **k-Anonymität**, wenn die Werte der Quasi-Identifikatoren eines jeden Individuums in diesem Datensatz mit den Werten der Quasi-Identifikatoren von mindestens k-1 anderen Individuen in diesem Datensatz übereinstimmen.
- Beispiel:

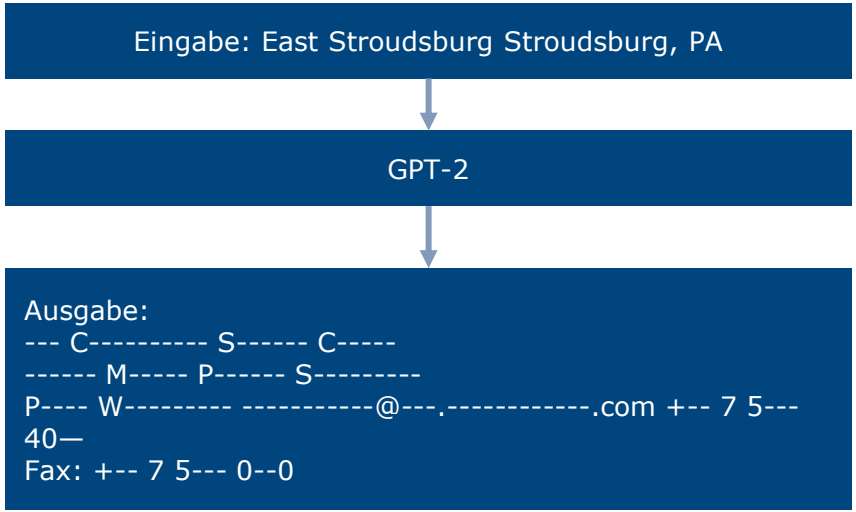
PLZ	Alter	Geschlecht	Krankheit	
81**	<60	M	Demenz	4-anonym
81**	<60	M	Demenz	
81**	<60	M	Herzkrankheit	
81**	<60	M	Herzkrankheit	
85**	≥60	M	Demenz	3-anonym
85**	≥60	M	Demenz	
85**	≥60	M	Demenz	

- Schwachstelle **Homogenitäts-Angriff**:
  - Innerhalb einer Äquivalenzklasse gibt es keine/wenige Unterschiede in den sensitiven Attributen.
  - Kennt man die Zugehörigkeit eines Individuums zu einer Äquivalenzklasse, dann kennt man den Wert des sensitiven Attributes.

Anonymität von Modellen

- Modelle können auch ein Datenschutzrisiko darstellen!
- Analog zu Daten gibt es Anonymisierungstechniken, -metriken, -angriffe.
- Populärstes Konzept ist **Differential Privacy**.

Beispiel aus dem Bereich Transformer-Modelle (aus N. Carlini et al. https://arxiv.org/pdf/2012.07805.pdf)



Inhalt

Ausblick

1. Veröffentlichung in der Zeitschrift „Der Aktuar“
   - 3. Ausgabe 2022
2. Erstellung von Schulungsunterlagen (ggf. Webinar)
   - Ende 2022
3. Ergänzung von Ausbildungsunterlagen
   - offen