Datenanonymisierung und Risikomanagement mit ARX

Dr. Fabian Prasser

Lehrstuhl für Medizinische Informatik
Institut für Medizinische Statistik und Epidemiologie
Klinikum rechts der Isar der TU München

fabian.prasser@tum.de
+49 89 4140 - 4328

Motivation und Hintergrund

Wesentliche Anwendungsfälle

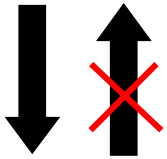
- Sekundärdatennutzung: Insbesondere Routinedaten für die Forschung
- Kollaborative Forschung: Data Sharing

Rolle des Datenschutzes

- Einhaltung rechtlicher Vorgaben
- Stärkung von Vertrauen, gesellschaftliche Akzeptanz

Dichotomie zwischen anonymen und personenbezogenen Daten

- Personenbezogene Daten



Wobei alle Mittel "die [...] wahrscheinlich genutzt werden" unter Bezug auf objektive Faktoren, wie "Kosten [...] Zeitaufwand, [...] verfügbare Technologie und technologische Entwicklungen" berücksichtigt werden müssen [1]

- Anonyme Daten

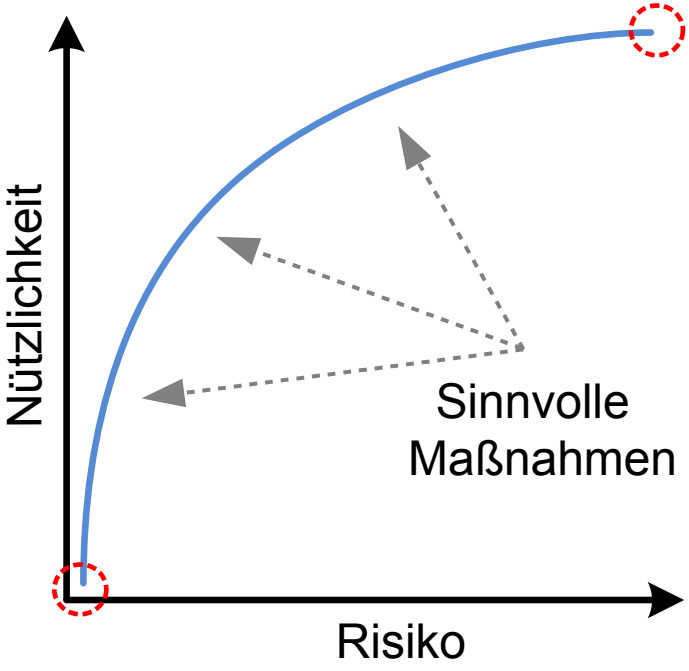
„Gesundheitsdaten“ gelten als besonders schützenswert [1]

[1] Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)

Herausforderungen

Grenze zwischen personenbezogenen und anonymen Daten ist fließend

- Datenschutz ist ein Abwägungsprozess
- Zentrale Herausforderung: Datenschutzrisiken vs. Nützlichkeit
- Ziel: starke Reduktion des Risikos bei geringer Auswirkung auf die Nützlichkeit
- Risikomanagement für Mikrodaten ist ein Baustein in einem Bündel von Maßnahmen



Anonymisierung

Originaldaten
Größtes Risiko

Vergröberung
der Daten

Keine Daten
Kein Risiko

Pseudonymisierung

Alle Daten in MDAT
Größtes Risiko

Verschieben
von Attributen
MDAT → IDAT

Keine Daten in MDAT
Kein Risiko

ARX

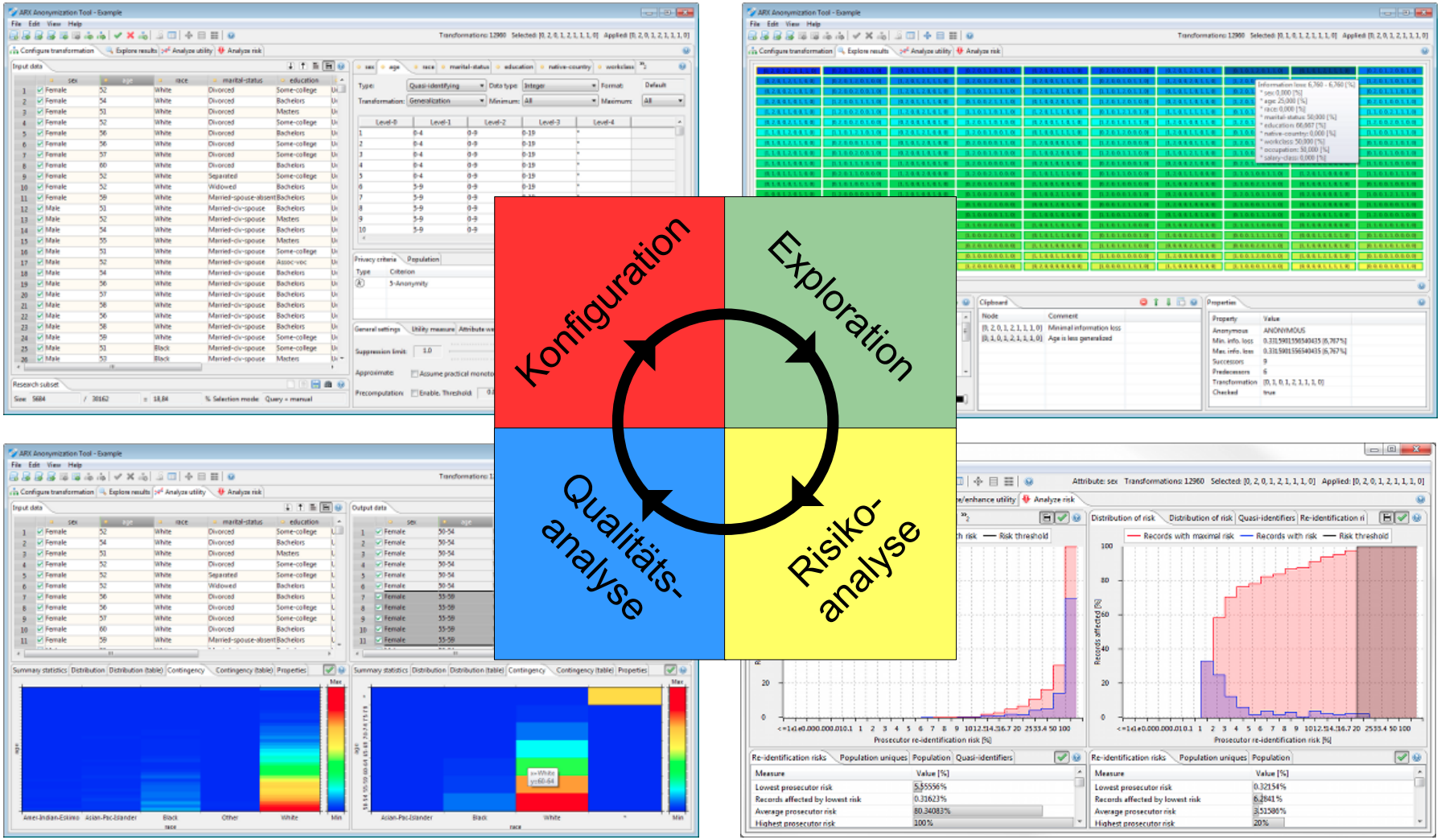
Hintergrund

- Software zur Analyse und Reduktion des Reidentifikationsrisikos strukturierter, tabellarischer Daten
- Methoden zur Messung und Reduktion von Risiken orientieren sich an Empfehlungen für die medizinische Domäne
- Für Datenanonymisierungslösungen besitzt die Software eine sehr intuitive graphische Benutzeroberfläche und einen hohen Grad an Automatisierung
- Schnittstellen zu gängigen Datenbanksystemen (MS SQL, DB2, SQLite, MySQL) und Tabellenkalkulationen (MS Excel, CSV)

Literaturempfehlungen (Auszug)

- George T. Duncan, Mark Elliot, Gonzalez Juan Jose Salazar. Statistical Confidentiality - Principles and Practice. Springer. 2011.
- Khaled El Emam. Guide to the De-Identification of Personal Health Information. CRC Press, 2013.

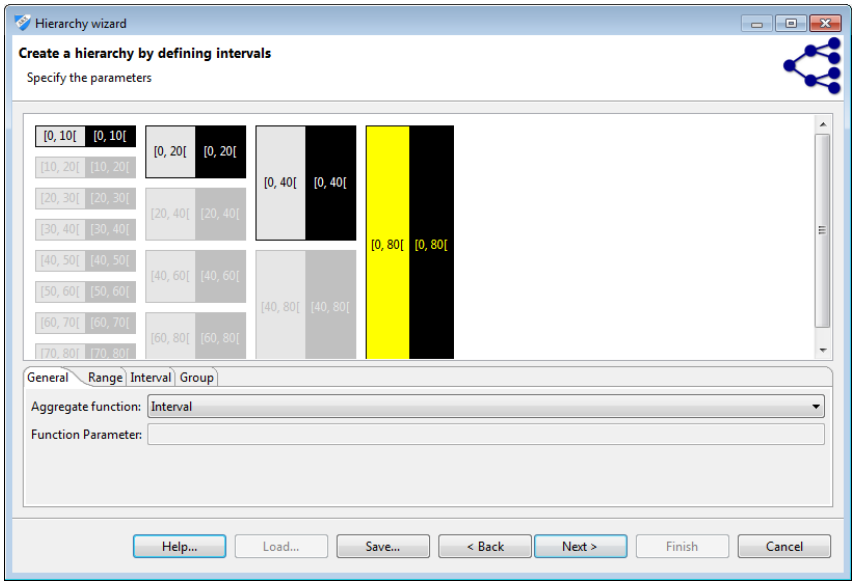
ARX: Perspektiven



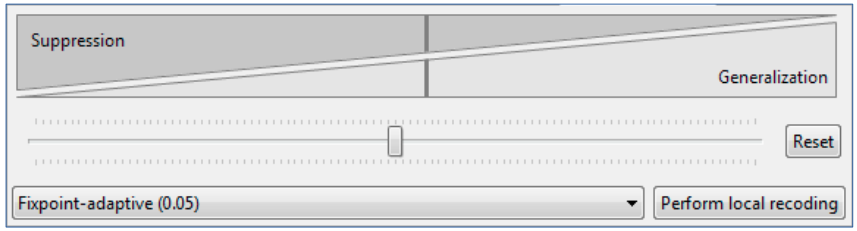
ARX: Konfiguration

Beispielsweise

- Risikogrenzwerte für verschiedene Arten von Angriffen
- Wichtigkeit einzelner Variablen
- Hintergrundwissen potentieller Angreifer
- Informationen zur Gesamtpopulation
- Transformationsmethoden und -regeln



sex	age	race	marital-status	education	native-coun...
Female	22	White	Married-civ-spouse	5th-6th	Mexico
Female	61	White	Married-civ-spouse	5th-6th	Mexico
Female	41	White	Married-civ-spouse	5th-6th	Mexico
Female	42	White	Married-civ-spouse	5th-6th	Mexico
Female	46	White	Married-civ-spouse	Preschool	Mexico
Female	36	White	Married-civ-spouse	Bachelors	Germany
Female	40	White	Married-civ-spouse	9th	Yugoslavia
Female	32	White	Married-civ-spouse	Some-college	France
Female	44	White	Married-civ-spouse	HS-grad	Italy
Female	60	White	Married-civ-spouse	Assoc-voc	Germany



ARX: Datentransformation

sex	age	race	marital-status	education	native-coun...
Female	22	White	Married-civ-spouse	5th-6th	Mexico
Female	61	White	Married-civ-spouse	5th-6th	Mexico
Female	41	White	Married-civ-spouse	5th-6th	Mexico
Female	42	White	Married-civ-spouse	5th-6th	Mexico
Female	46	White	Married-civ-spouse	Preschool	Mexico
Female	36	White	Married-civ-spouse	Bachelors	Germany
Female	40	White	Married-civ-spouse	9th	Yugoslavia
Female	32	White	Married-civ-spouse	Some-college	France
Female	44	White	Married-civ-spouse	HS-grad	Italy
Female	60	White	Married-civ-spouse	Assoc-voc	Germany

Reduktion der Granularität potentiell identifizierender Variablen

sex	age	race	marital-status	education	native-coun...
Female	40	White	Married-civ-spouse	Primary School	Mexico
Female	40	White	Married-civ-spouse	Primary School	Mexico
Female	40	White	Married-civ-spouse	Primary School	Mexico
Female	40	White	Married-civ-spouse	Primary School	Mexico
Female	40	White	Married-civ-spouse	Primary School	Mexico
Female	41	White	Married-civ-spouse	*	Europe
Female	41	White	Married-civ-spouse	*	Europe
Female	41	White	Married-civ-spouse	*	Europe
Female	41	White	Married-civ-spouse	*	Europe
Female	41	White	Married-civ-spouse	*	Europe

Stichprobe

Generalisierung

Unterdrückung

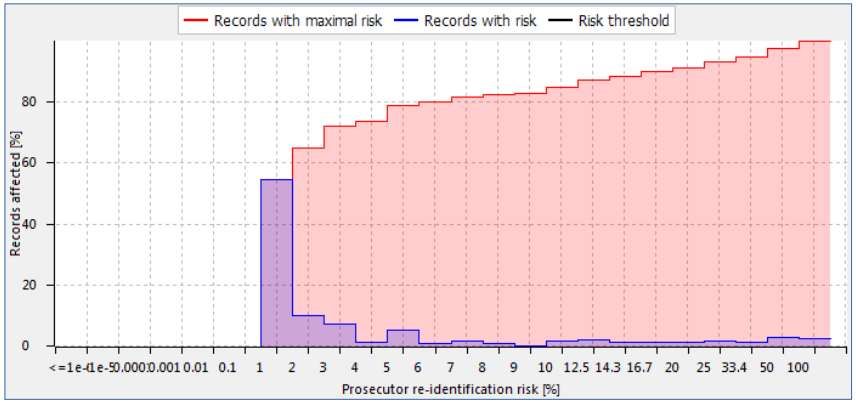
Mikroaggregation

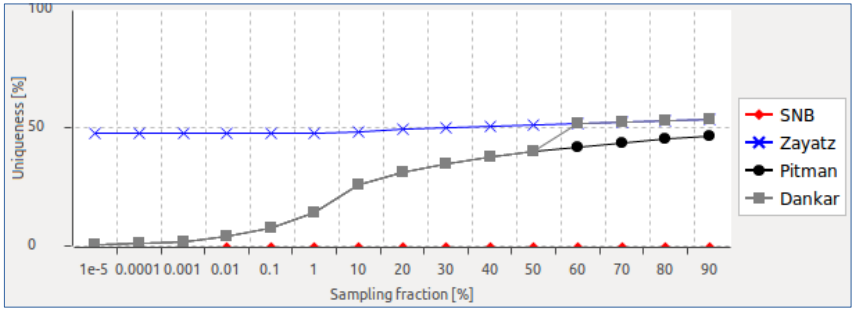
Top-/Bottom Coding

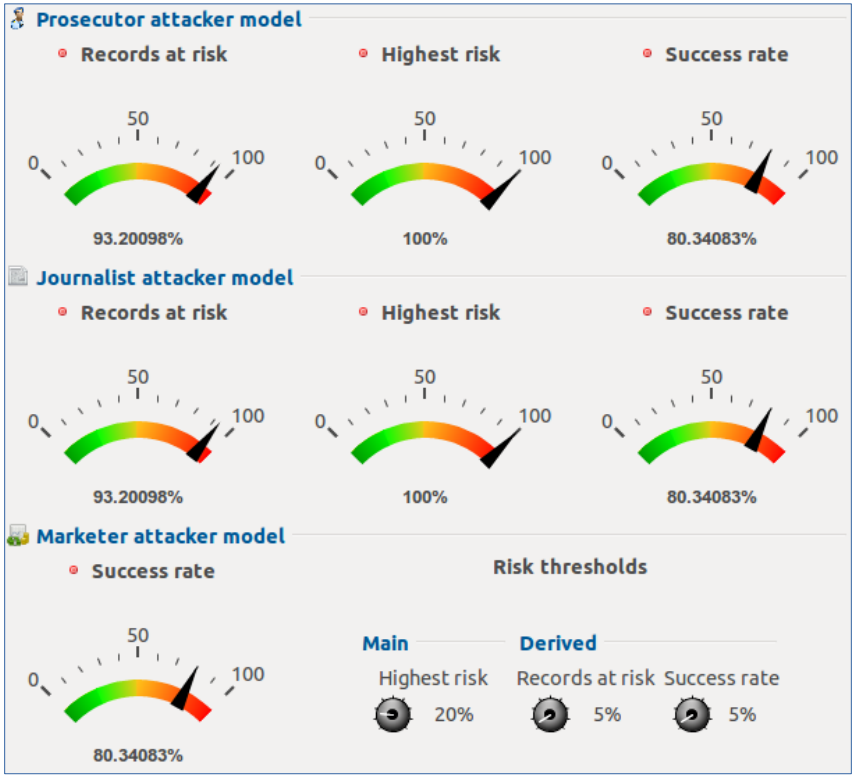
ARX: Risikoanalyse

Verschiedene Modelle

- Besonderer Fokus auf Re-Identifikation
- Vorher/nachher Vergleiche



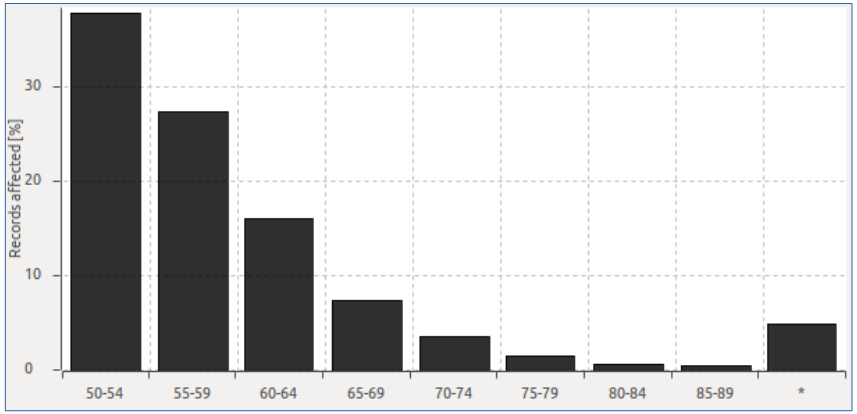


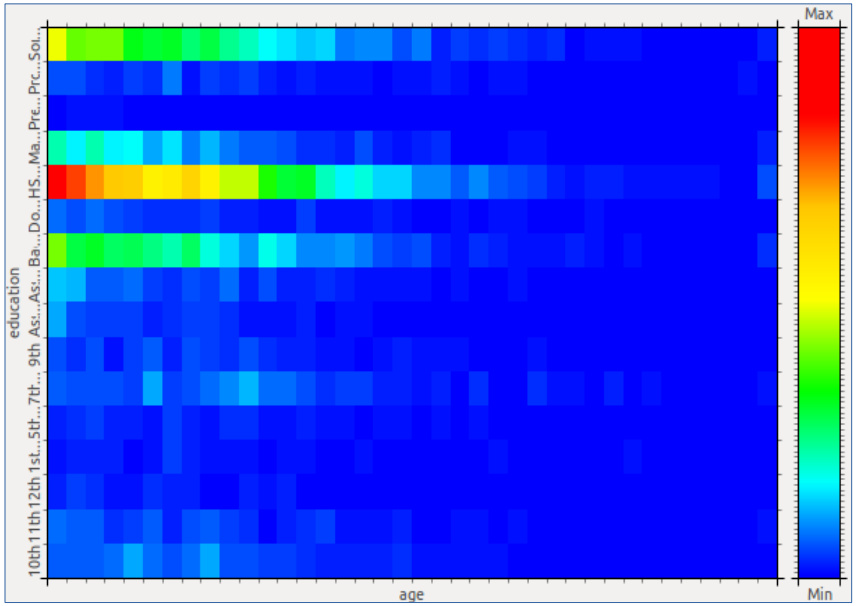


ARX: Qualitätsanalyse

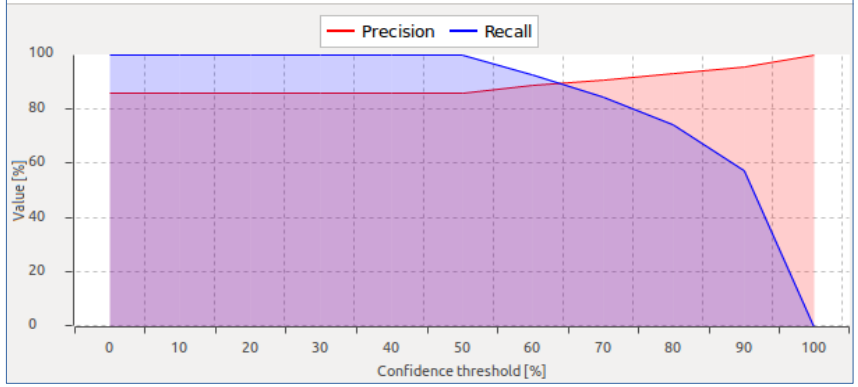
Mehrere Verfahren

- Modelle für den Informationsgehalt
- Deskriptive Statistik
- Statistische Klassifikation
- Vorher/nachher Vergleiche





Class	Instances	Baseline accu	Accuracy	Gain/loss	Avg. error
education	16	34.17077%	36.80604%	2.63528%	76.51772%
race	5	88.15882%	89.35348%	1.19466%	18.71493%
sex	2	70.90654%	85.94519%	15.03865%	19.98407%
marital-status	6	62.29796%	70.95924%	8.66128%	42.5385%
age	38	10.03162%	7.51933%	-2.5123%	94.40756%



ARX: Besonderheiten

Umfangreiche Methodenunterstützung

- Beispiel: 15 verschiedene Risiko-/Datenschutzmodelle
- Verfügbar für alle wesentlichen Betriebssysteme (Windows, MacOS, Linux)
- Graphisches Werkzeug und Programmierbibliothek

Hochskalierbar

- Verarbeitung mehrerer Millionen Datensätze mit bis zu 50 potentiell identifzierbaren Variablen auf Consumer Hardware

International erfolgreich

- >20.000 Downloads seit 2012
- EMA: External Guidance on the Implementation of the European Medicines Agency Policy 0070 on the Publication of Clinical Data for Medicinal Products for Human Use
- EU Agency for Network and Information Security (ENISA): Privacy and Data Protection by Design
- NIST Special Publication 800-188 (Draft): De-Identifying Government Datasets

Open Source Software

- Lizenz: Apache 2

Danke für Ihre Aufmerksamkeit!

Dr. rer. nat. Fabian Prasser

Klinikum rechts der Isar
Technische Universität München
Institut für Medizinische
Statistik und Epidemiologie

Ismaninger Str. 22
81675 Munich
Germany

Tel +49 89 4140-4328
Fax +49 89 4140-4850
fabian.prasser@tum.de
www.imse.med.tum.de
arx.deidentifier.org