Technischer Datenschutz im KI-Kontext

Hannes Federrath
Sicherheit in verteilten Systemen (SVS)
http://svs.informatik.uni-hamburg.de

Der Arbeitsbereich Sicherheit in Verteilten Systemen (SVS)

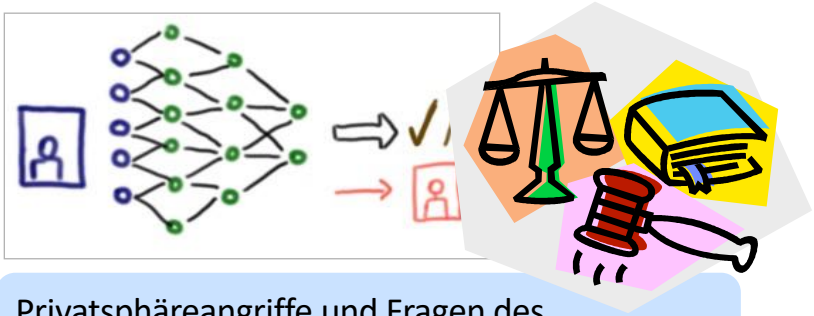
- Unsere Forschungsthemen (Auswahl)
  - IT-Sicherheitsmanagement und -Grundschutz, ISO 27001
  - Privacy im Internet, Schutz vor Beobachtung, IT-Forensik
  - Sichere und datenschutzfreundliche Vernetzung von Fahrzeugen
  - Sicherheit und Datenschutz in mobilen Systemen

- Beiträge und (interdisziplinäre) Ergebnisse
  - Begleitung von Gesetzgebungsverfahren aus technischer Sicht
  - Erforschung des Spannungsfeldes von Freiheit und Sicherheit
  - Technische Lösungen zum Grundrechtsschutz
  - Informatik als gesellschaftliche Aufgabe

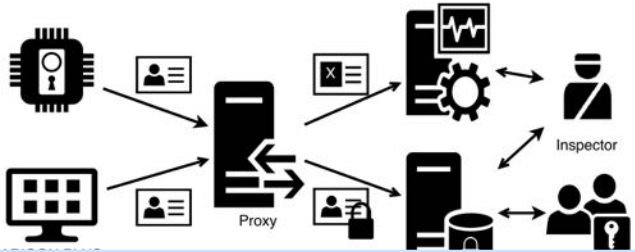
- Weitere Informationen
  - https://svs.informatik.uni-hamburg.de

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

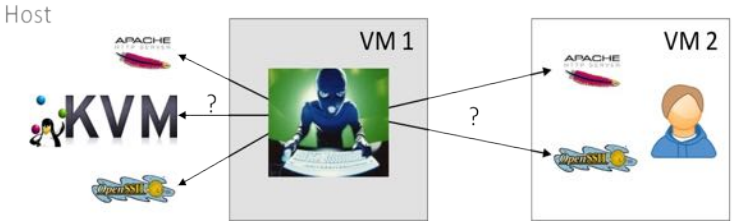
Forschung: Datenschutz und Datensicherheit (Auswahl)



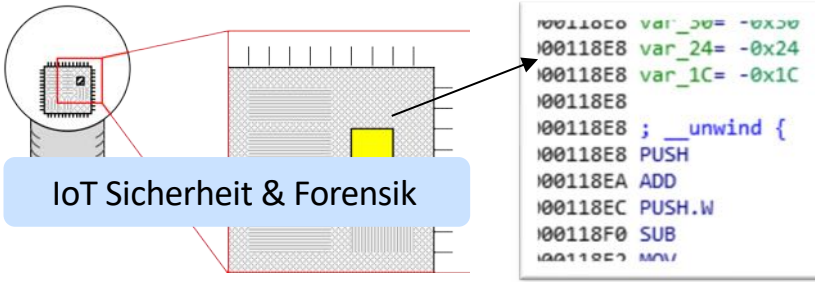
Privatsphäreangriffe und Fragen des technischen Datenschutzes in der KI



Sicherheit und Privatsphäre in Anwendungen, Datenschutzfreundliche Angriffserkennung



Sicherheit und Privatsphäre in der Cloud



IoT Sicherheit & Forensik

Datenschutz

Datenschutz dient nicht nur

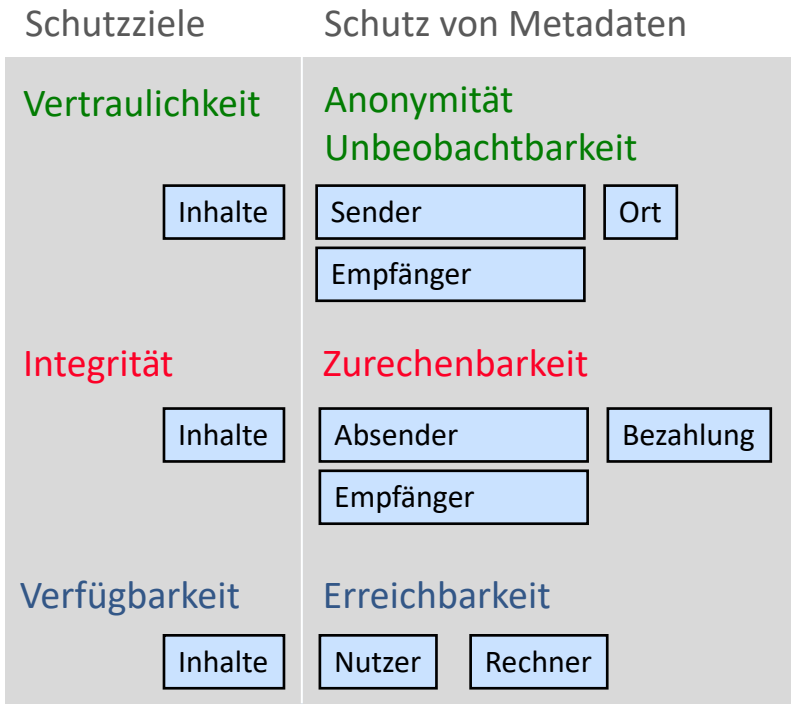
- dem Schutz vor unberechtigter Kenntnisnahme persönlicher Daten (Schutzziel Vertraulichkeit),

sondern auch

- dem Schutz vor absichtlicher oder versehentlicher Verfälschung und missbräuchlicher Verwendung (Schutzziel Integrität) sowie
- dem Schutz vor Verlust von persönlichen Daten (Schutzziel Verfügbarkeit).

Verletzlichkeiten führen selten nur zur Verletzung eines Schutzziels, sondern führen bei einem Vorfall meist zu Schäden sowohl der Vertraulichkeit von Daten als auch der Integrität und Verfügbarkeit.

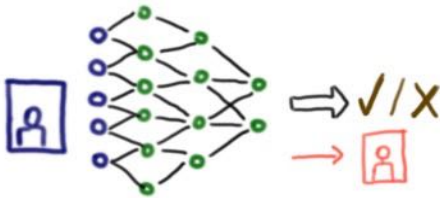
–> Fallbeispiel: Identitätsdiebstahl



Beachte: Datenschutz = Schutz der Menschen
Schutz der Daten = Datensicherheit

Thesen zum Datenschutz im KI-Kontext

- Training von KI-Algorithmen benötigt große Mengen an Daten
  - oftmals personenbezogene Daten
  - medizinischer Bereich: sensible Daten (Art. 9 DSGVO)
- Zensieren von sensiblen Daten reicht nicht aus
  - von KI reproduzierbar
- KI-Modelle lernen mehr als für die jeweilige Aufgabe notwendig
  - Repräsentation der Daten bleibt in KI-Modell bestehen
    - auch von einzelnen Datenpunkten



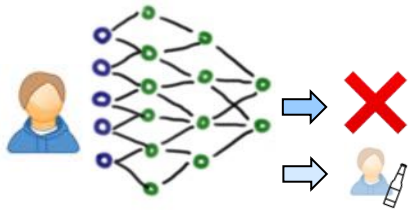
Folge: Privatsphäre-Angriffe sind u.U. möglich. Die Abschätzung der Risiken *vor* der Einführung eines (KI)-Verfahrens ist notwendig.

Beispiel 1: Membership Inference Angriff

Shokri, Stronati, Song, Shmatikov, 2017

- Patient Max: alkoholkrank
- Einwilligung in Suchtklinik zu Verarbeitung seiner anonymisierten Patientendaten
  - Training eines KI-Modells der Firma AlcScore

- Idee von AlcScore:
  - Anhand von Patientendaten Risikofaktoren extrahieren
  - Eingabe in KI-Modell: Anonymisierte Patientendaten
  - Ausgabe: Risikoscore Neigung zu Alkoholismus

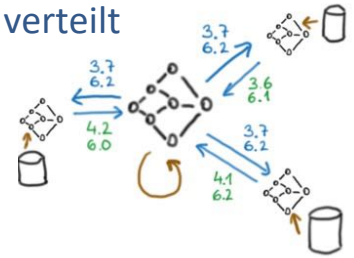


- Jahre später: Max möchte Lebensversicherung abschließen und muss Gesundheitsdaten angeben
  - Versicherung hat Modell von AlcScore akquiriert
  - Kann anhand Max‘ Daten nicht nur Score berechnen
  - durch Membership Inference: Max war Teil der AlcScore-Trainingsdaten
  - Diagnostizierte Sucht –> Versicherung lehnt Vertrag mit Max ab

Beispiel 2: Federated Learning Rekonstruktionsangriff

Zhu, Han, 2020

- Federated Learning (FL): Trainingsdaten für Modell sind auf mehrere Standorte verteilt
  - Vorteil: Daten müssen nicht ausgetauscht werden
  - Zentraler Server koordiniert
    - Jeder Standort trainiert lokales Modell
    - Schickt aktualisierte Modellparameter an Server
    - Server aggregiert Parameter, aktualisiert und verteilt globales Modell



Definition Federated Learning

Federated Learning beschreibt ein Szenario des maschinellen Lernens (ML), in dem mehrere Entitäten (Clients) kooperativ und unter Koordinierung durch eine zentrale Instanz (Server oder Service Provider) ein ML-Problem lösen.

Die Daten der Clients werden lokal gespeichert und nicht ausgetauscht oder übertragen, stattdessen wird das Lernziel durch gezielte Modellupdates und ihre unmittelbar erfolgende Aggregation erreicht.

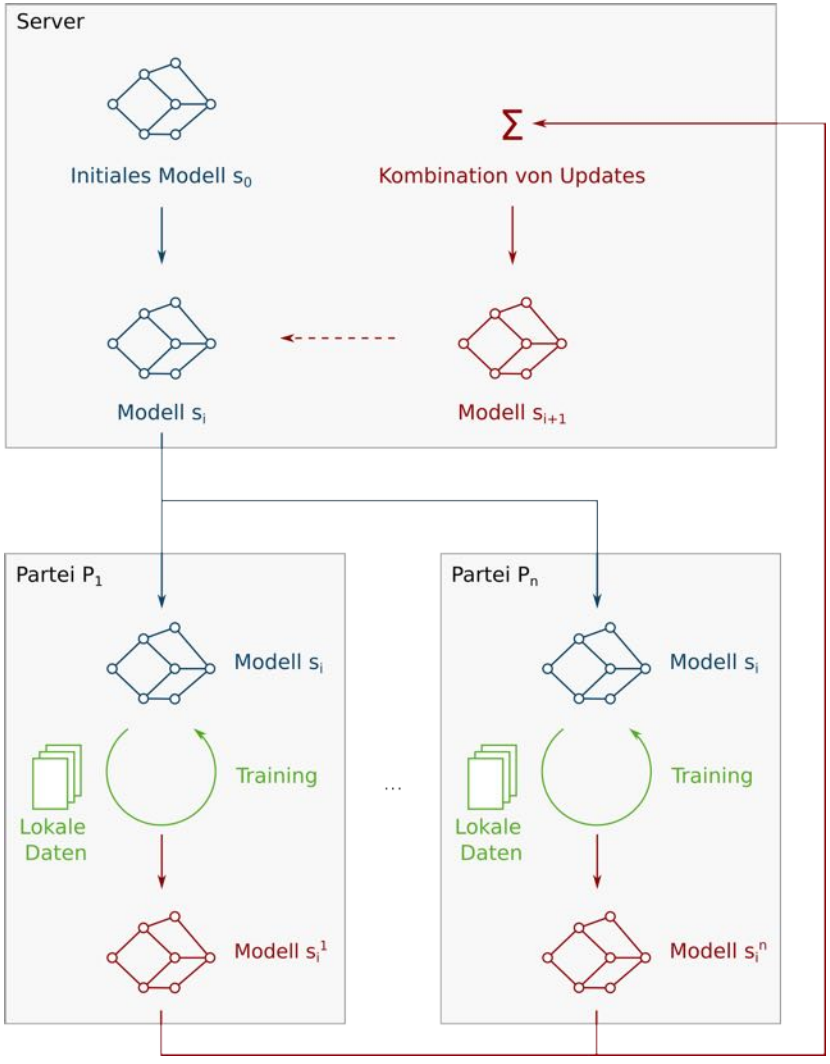
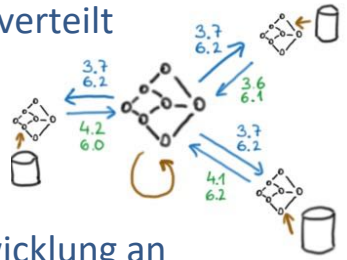


Abb.: Stock et al. »Privatsphärefreundliches maschinelles Lernen«, Informatik Spektrum 45/2 (2022)

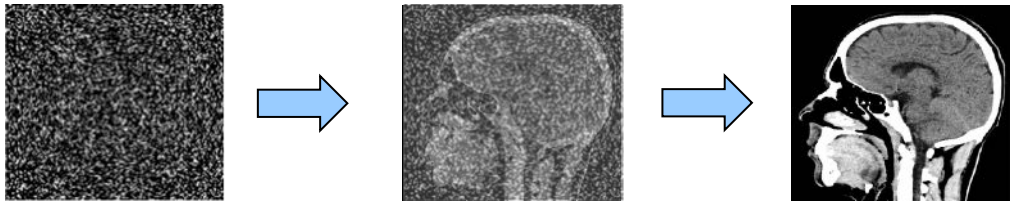
Beispiel 2: Federated Learning Rekonstruktionsangriff

Zhu, Han, 2020

- Federated Learning (FL): Trainingsdaten für Modell sind auf mehrere Standorte verteilt
  - Vorteil: Daten müssen nicht ausgetauscht werden
  - Zentraler Server koordiniert, aggregiert und verteilt globales Modell



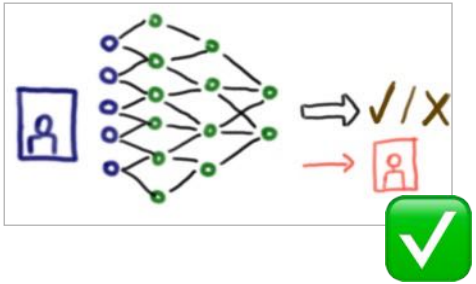
- Beispiel: Gesundheitsdienstleister bietet FL-Dienst für gemeinsame Modellentwicklung an
  - Hirntumorerkennung auf CT-Bildern
  - Kliniken nehmen teil: lediglich Modellupdates, aber keine CT-Bilder werden übertragen
  - Dienstleister kann aus Modellupdates einzelner Klinik Bilder rekonstruieren
    - Start bei zufälligem Rauschen
    - iterativer Prozess, bis Modellparameter 1:1 übereinstimmen



Lösungswege Privacy Preserving Machine Learning (PPML)

- Anonymisierung und Pseudonymisierung
  - Nicht trivial; bedingt wirksam gegen KI-Privatsphäreangriffe
- Differential Privacy
  - Kontextabhängige, gezielte Manipulation von Algorithmen oder Daten
  - Minimiert Einfluss einzelner Datenpunkte –> wirksam gegen Membership Inference und Rekonstruktionsangriffe
- Secure Multiparty Computation (SMPC)
  - Kryptographische Lösung für verteiltes Training
- Homomorphe Verschlüsselung (HE)
  - Ermöglicht Training und Inferenz auf verschlüsselten Daten

  - SMPC und HE: rechenintensiv und (noch) nicht für Masseneinsatz tauglich



Privacy by Design
Art. 25 und 32 DSGVO

Privacy by Design

Stand der Technik

Kosten

ROSI

Art, Umfang, Umstände, Zwecke

Privacy by Default

- Beschränkung auf Verarbeitungszweck
- Beschränkung des Umfangs
- Beschränkung auf Speicherfristen
- Beschränkung der Zugänglichkeit

Risiko-management

Technisch-organisatorische Maßnahmen

Datenschutzgrundsätze

Betroffenenrechte

Rechtmäßigkeit, Transparenz

Zweck-bindung

Rechenschafts-pflicht

Richtigkeit

Speicher-begrenzung

Daten-minimierung

Daten-sicherheit

- Auskunft
- Berichtigung
- Löschung
- Einschränkung der Verarbeitung
- Datenübertragbarkeit
- Widerspruchsrecht

- Verbot mit Erlaubnisvorbehalt oder Einwilligung
- Ersterhebung bei betroffener Person
- Gute Dokumentation
- ...

Privacy Policy

- Anonymität
- Unbeobachtbarkeit
- Unverkettbarkeit

- Pseudonymisierung
- Verschlüsselung
- Vertraulichkeit
- Integrität
- Verfügbarkeit
- Belastbarkeit
- Wiederherstellbarkeit
- Notfallmanagement
- Überprüfbarkeit
- Bewertung
- Evaluierung

PIMS

Bereitschaft zur datenschutz-freundlichen Gestaltung

Big Data

Algorithmen

- Enforce
- Inform
- Control
- Demonstrate

- Minimise
- Separate
- Aggregate
- Perturbate
- Hide

Initiierung

Si-Konzept

Umsetzung

Erhaltung

Kontext

Bewertung

Gew-Ziele

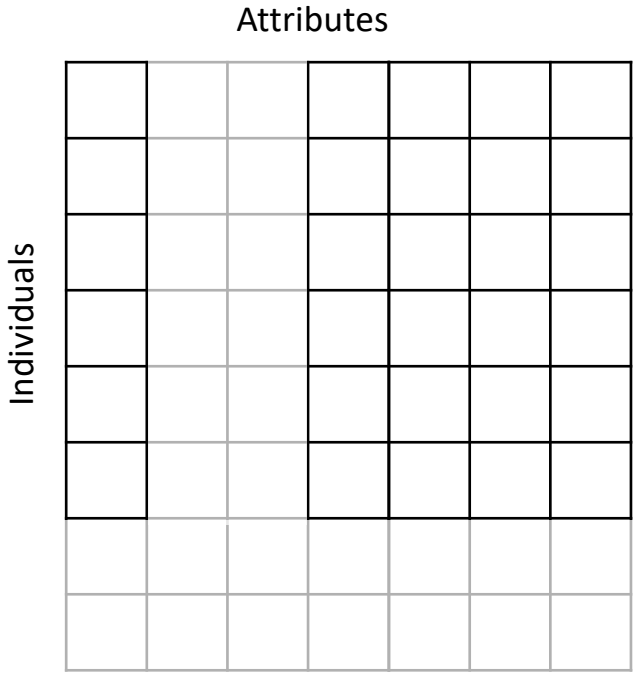
Soll-Ist-An

Feedback

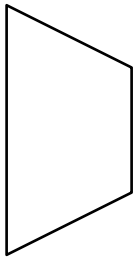
Bereitschaft

Privacy design strategies

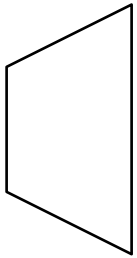
in Anlehnung an Hoepman, 2014



minimise

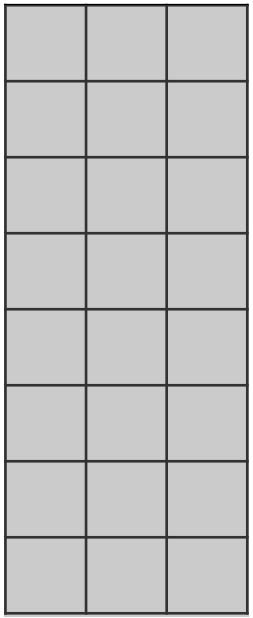


aggregate



perturbate

separate



hide

Privacy by Design (Art. 25 DSGVO) und Sicherheit der Verarbeitung (Art. 32 DSGVO)

Minimise: Nur notwendige Daten speichern und verarbeiten	Anonymisierung, Pseudonymisierung
Separate: Daten verteilt verarbeiten und speichern	Federated Learning, SMPC
Aggregate: Daten auf das notwendige Maß zusammenfassen	Federated Learning
Perturbate: Daten durch zufällige Störungen ungenau machen	Differential Privacy
Hide: Daten nicht in offener Form speichern	Homomorphic Encryption

Pseudonymisierte Daten

- »Einer Pseudonymisierung unterzogene personenbezogene Daten, die durch Heranziehung zusätzlicher Informationen einer natürlichen Person zugeordnet werden könnten, sollten als Informationen über eine identifizierbare natürliche Person betrachtet werden.« (aus: ErwG 26)
  - Getrennte Datenhaltung
  - Explizite Zuordnungsregel vorhanden

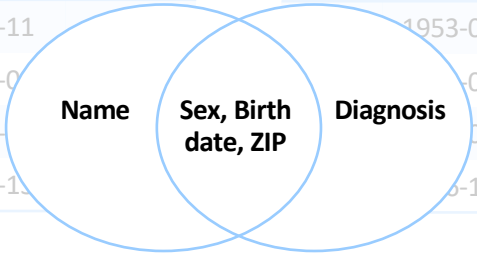
Name	P
Alice	735
Bob	324
Carol	478
Dan	125

P	Sex	Diagnosis
735	f	A
324	m	B
478	f	C
125	m	D

- Auch Merkmalskombinationen stellen ein Pseudonym dar, wenn sie Eindeutigkeit erzielen.
  - Getrennte Datenhaltung
  - Fehlen einer expliziten Zuordnungsregel
  - Quasi-Identifiers (QIDs)

Name	Sex	Birth date	ZIP
Alice	f	1953-06-11	
Bob	m	1922-03-0	
Carol	f	1973-05-	
Dan	m	1966-10-1	

Sex	Birth date	ZIP	Diagnosis
	1953-06-11	12345	A
	-03-02	98763	B
	05-20	12390	C
	-10-13	98764	D



Anonymisierte Daten

- Vermeintlich einfache Umsetzbarkeit von Anonymisierung
  - Entfernen der Datenfelder mit Personenbezug
- Probleme
  - fehlende Entscheidbarkeit, ob verbleibende Felder Quasi-Identifiers (QIDs)
  - Kontextwissen eines Angreifers ist zum Zeitpunkt der Anonymisierung unbekannt

- Metriken zur Messung des Anonymisierungsgrades
  - k-Anonymität (Sweeney, 2002)
    - Für alle Datensätze gilt, dass nach der Anonymisierung wenigstens k Datensätze nicht mehr voneinander unterscheidbar sind.

    Generalisieren
    Aggregieren
  - Differential Privacy (Dwork, 2006)
    - Für alle Datensätze, die sich in höchstens einem Eintrag unterscheiden, ist die Wahrscheinlichkeit kleiner als ein vorgegebener Wert e^{ε}, dass diese nach der Anonymisierung noch unterscheidbar sind.

    Perturbieren
    Verrauschen
  - … (zahlreiche weitere)

k-Anonymität

Sweeney, 2002

- Jeder Datensatz ist bezüglich einer vorgegebenen Kombination von QIDs ununterscheidbar von wenigstens k-1 anderen Datensätzen.

Name	Sex	ZIP	Diagnosis
Alice	f	12345	A
Bob	m	98763	B
Carol	f	12390	C
Dan	m	98764	D
Eve	f	98765	E

k = 2

Name	Sex	ZIP	Diagnosis
	f	123**	A
	*	9876*	B
	f	123**	C
	*	9876*	D
	*	9876*	E

Privacy by Design

Stand der Technik

Kosten

- ROSI

Art, Umfang, Umstände, Zwecke

Privacy by Default

- Beschränkung auf Verarbeitungszweck
- Beschränkung des Umfangs
- Beschränkung auf Speicherfristen
- Beschränkung der Zugänglichkeit

Risiko-management

Technisch-organisatorische Maßnahmen

Datenschutzgrundsätze

Betroffenenrechte

- Auskunft
- Berichtigung
- Löschung
- Einschränkung der Verarbeitung
- Datenübertragbarkeit
- Widerspruchsrecht

Bereitschaft

Rechtmäßigkeit, Transparenz

Zweck-bindung

Rechenschafts-pflicht

Richtigkeit

Speicher-begrenzung

Daten-minimierung

Daten-sicherheit

- Verbot mit Erlaubnisvorbehalt oder Einwilligung
- Ersterhebung bei betroffener Person
- Gute Dokumentation
- ...

Privacy Policy

Bereitschaft zur datenschutz-freundlichen Gestaltung

- Anonymität
- Unbeobachtbarkeit
- Unverkettbarkeit

- Pseudonymisierung
- Verschlüsselung
- Vertraulichkeit
- Integrität
- Verfügbarkeit
- Belastbarkeit
- Wiederherstellbarkeit
- Notfallmanagement
- Überprüfbarkeit
- Bewertung
- Evaluierung

PIMS

Big Data und KI

Algorithmen

- Enforce
- Inform
- Control
- Demonstrate

- Minimise
- Separate
- Aggregate
- Perturbate
- Hide

Initiierung

Si-Konzept

Umsetzung

Erhaltung

Kontext

Bewertung

Gew-Ziele

Soll-Ist-An

Feedback

Extraktion von Eigenschaften aus Machine-Learning-Modellen

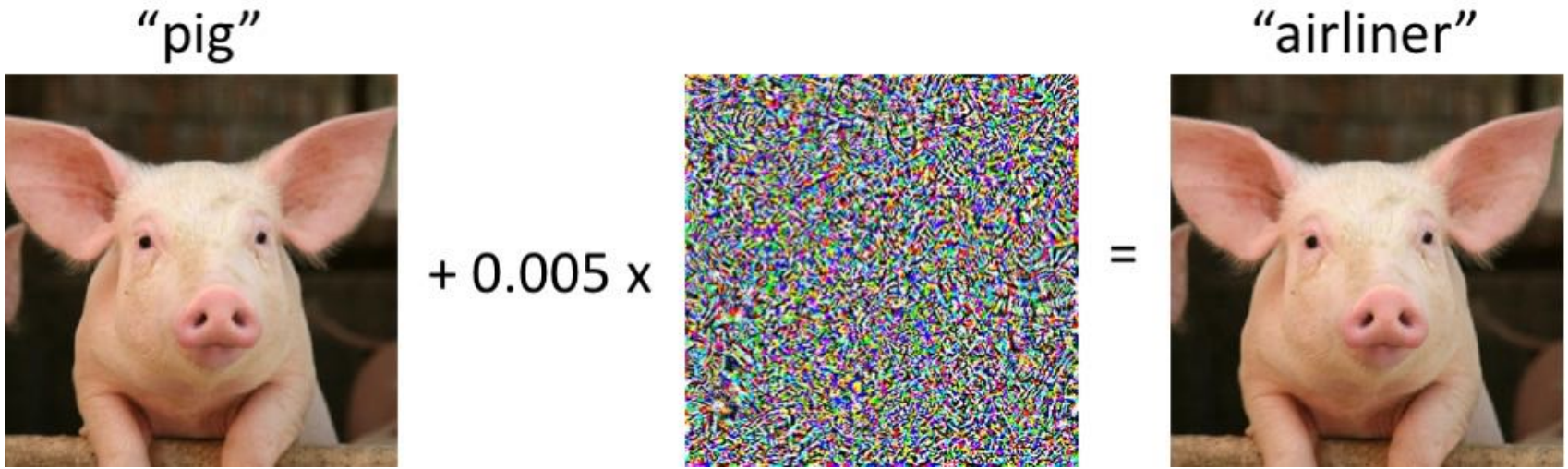
- These: personenbezogene Daten in Machine-Learning-Modellen sind aufgrund von Aggregation und damit verbundenem Informationsverlust nicht mehr rückführbar auf einzelne Individuen
  - Beispiel: Daten für das Anlernen KI-Systemen von stammen aus verschiedenen Quellen mit einheitlichen Merkmalen
    - Model M_A aus Quelle A: Trainingsdaten bestehen überwiegend aus Männern (70%)
    - Model M_B aus Quelle B: Trainingsdaten enthalten überwiegend Frauen (70%)
  - Angriffsziel: Separieren von M_A und M_B
- EuGH, Az. C-582/14 : Daten sind auch dann personenbezogen, wenn die Zuordnung zu einer Person nur indirekt vorgenommen werden kann.
  - Zuordnung über Gruppenzugehörigkeit genügt bereits
- Aktuelle Forschungsarbeiten zum sog. Property unlearning

Shokri, Reza, et al. Membership inference attacks against machine learning models. IEEE Symposium on Security and Privacy (SP) 2017. IEEE, 2017.

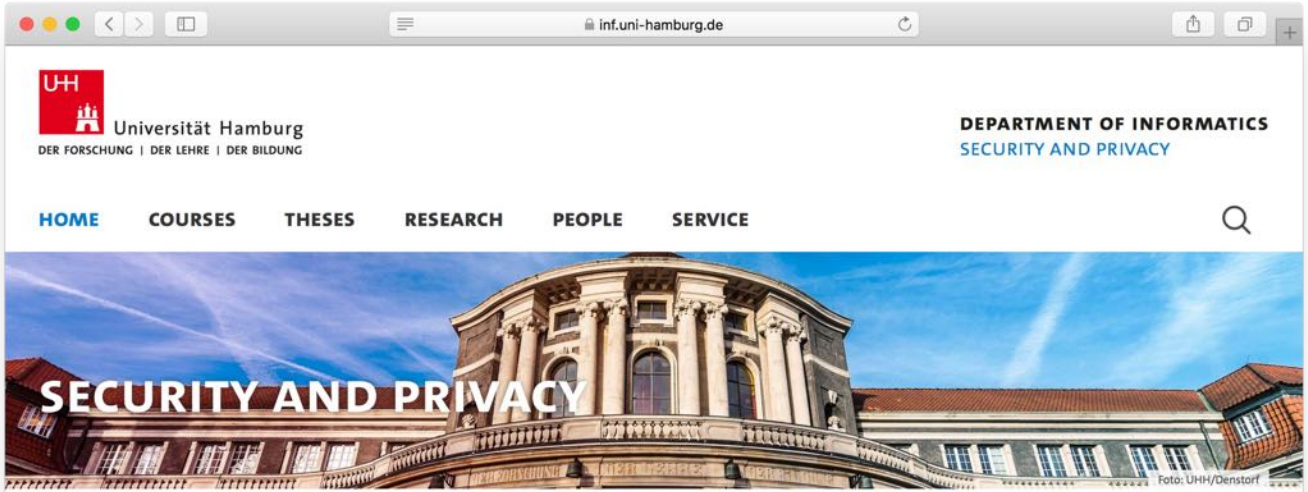
Zugespitzte Frage: War ein bestimmtes Individuum in das Training Data Set eingeschlossen?

Verbergen von Eigenschaften beim Klassifizieren

- Adversarial learning



https://www.designnews.com/electronics-test/yes-ai-can-be-tricked-and-its-serious-problem/161652909959780



Prof. Dr. Hannes Federrath
Fachbereich Informatik
Universität Hamburg
Vogt-Kölln-Straße 30
D-22527 Hamburg

Telefon +49 40 42883 2358

federrath@informatik.uni-hamburg.de

https://svs.informatik.uni-hamburg.de