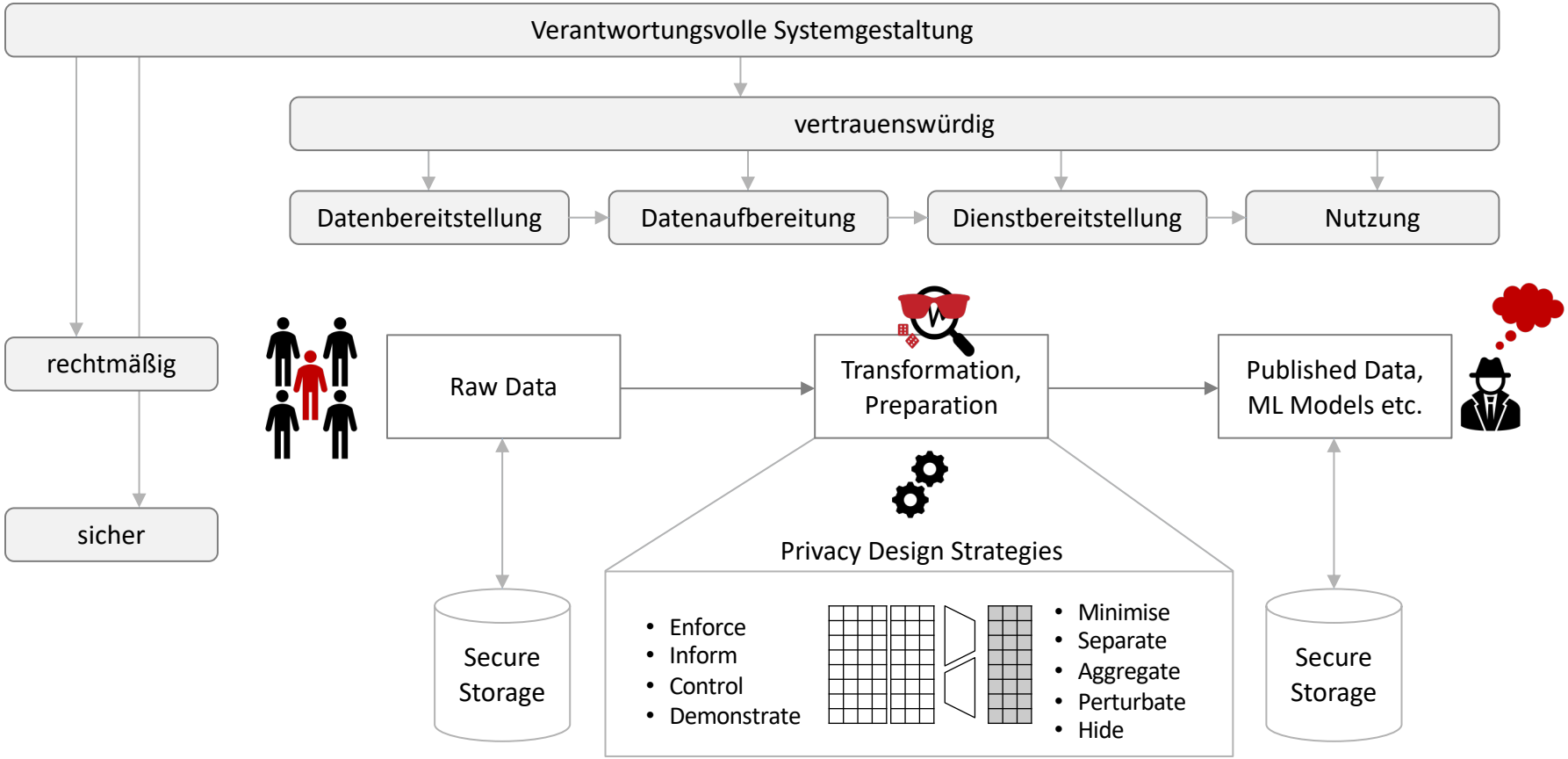
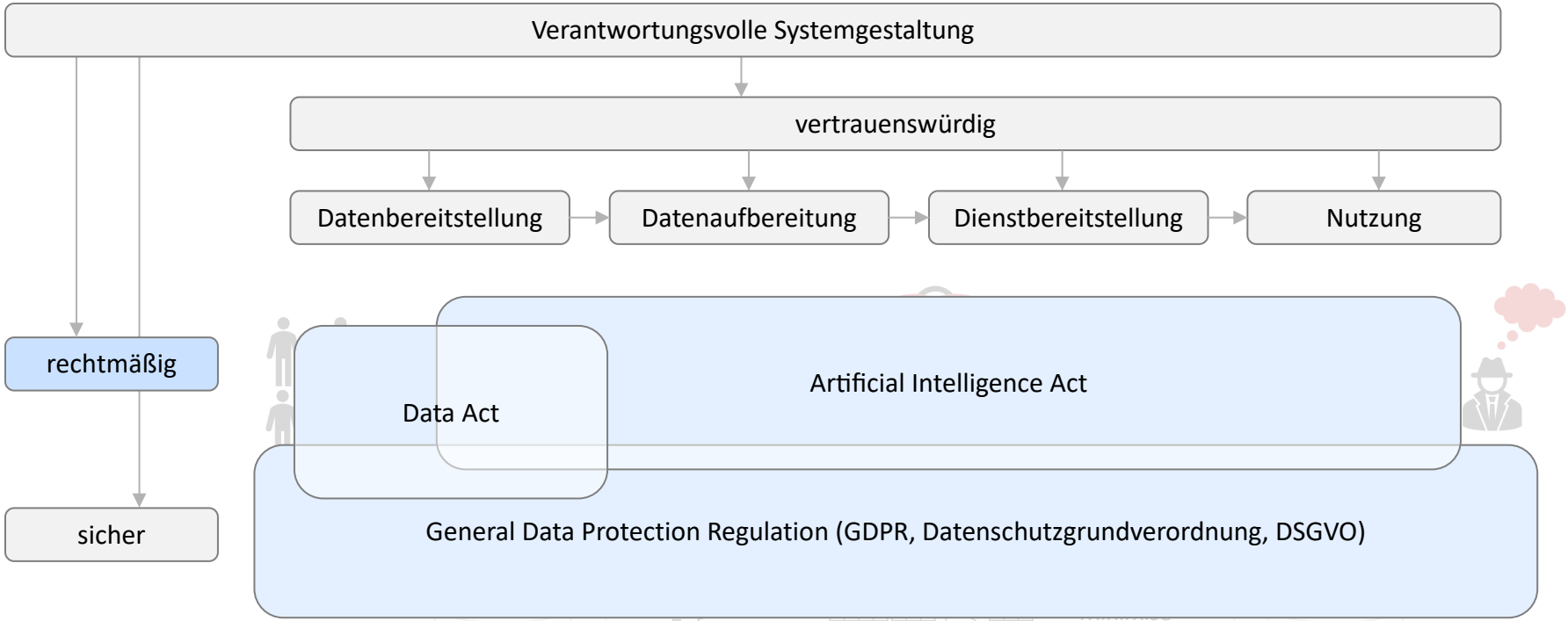
Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten

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

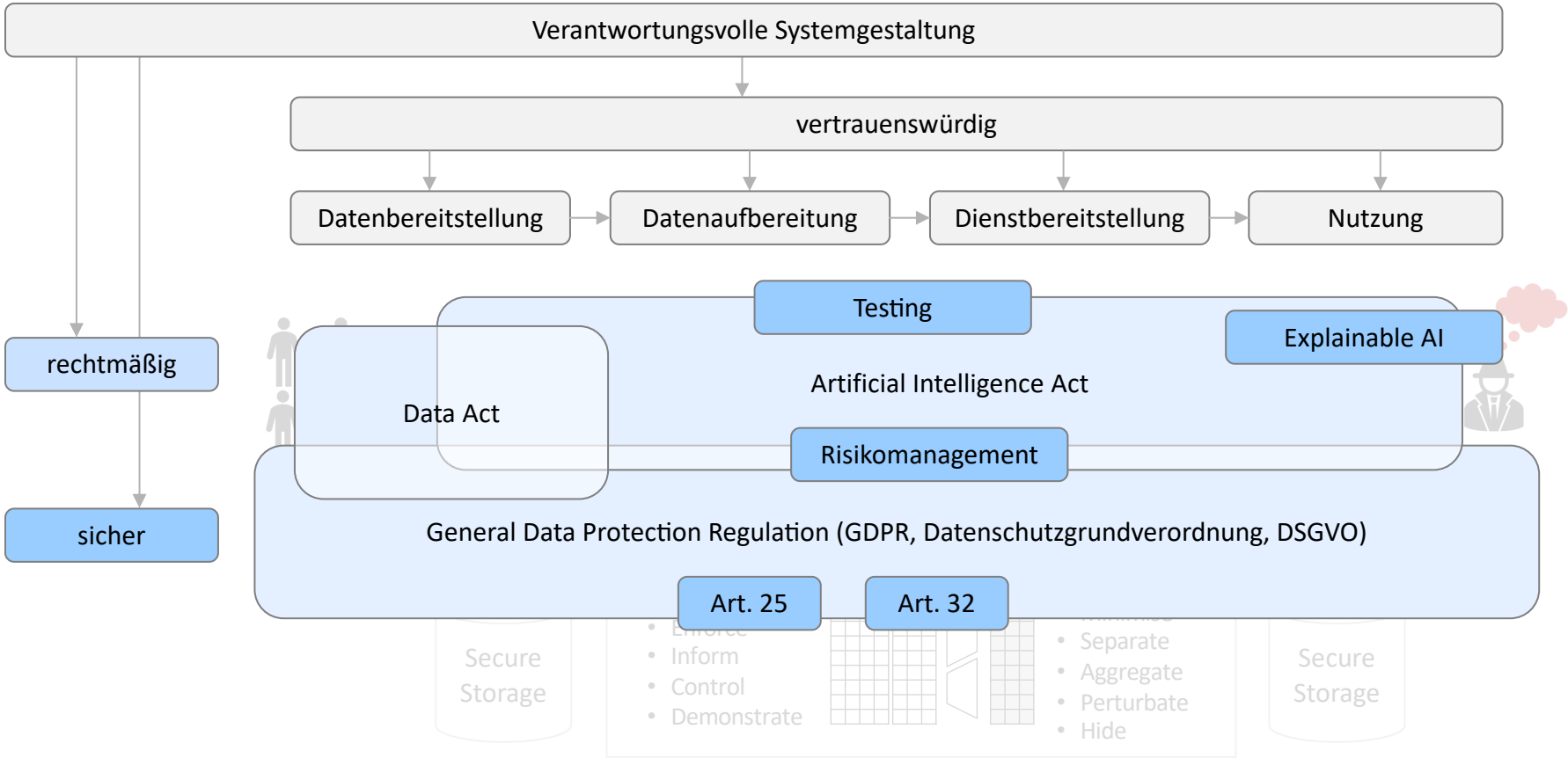
Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten



Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten



Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten



http://data.europa.eu/eli/reg/2024/1689/oj

The Artificial Intelligence Act (AI Act)

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

The Artificial Intelligence Act (AI Act)

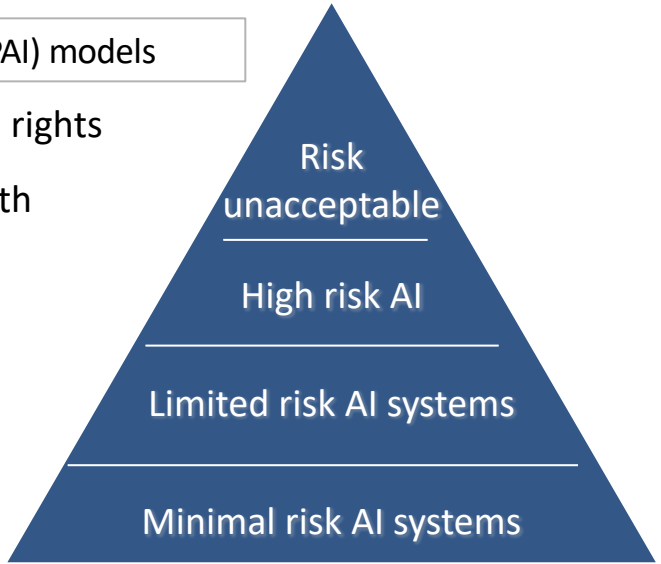
- Art. 3(1) Definition AI system

> 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments

- Risk-based classification approach

+ General Purpose AI (GPAI) models

  - Prohibited AI systems: contradict Union values, fundamental rights
  - High risk AI systems: Permitted but subject to compliance with specific product requirements and operator obligations
  - Limited risk AI systems: Permitted but subject to specific transparency and disclosure obligations
  - Minimal risk AI systems: Permitted, with no additional AI Act requirements. It is important to emphasize the importance of the GDPR in this context.



Timeline of the AI Act

- Phased approach over 3 years
  - 12 Jul 2024 Published
  - 1 Aug 2024 Entry into force (Art. 113, 20 days after publication)
  - 2 Feb 2025 General provisions & prohibited AI practices
  - 2 Aug 2025 Notifying authorities, General Purpose AI (GPAI) models, governance, penalties
  - 2 Aug 2026 General application of AI Act
  - 2 Aug 2027 Products covered by Union harmonisation legislation and corresponding obligations
- Contents
  - 180 recitals
  - 113 Articles
  - 13 Annexes
- Text
  - http://data.europa.eu/eli/reg/2024/1689/oj

Important note: The AI Act applies even when no personal data is processed – it is concerned with the broader ethical and safety implications of AI technologies.

General Purpose AI (GPAI) models

- Art. 3(1) Definition General Purpose AI

> 'general-purpose AI model' … trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications …

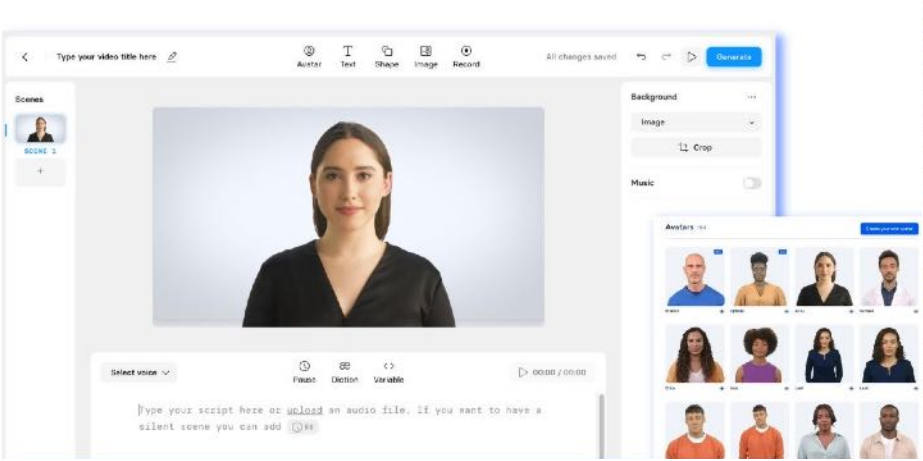
- GPAI models are treated as a separate category with their own specific regulations
  - Standard GPAI Models – subject to certain baseline obligations
  - GPAI Models with Systemic Risk (Art. 51) – more stringent requirements due to their potential for widespread impact.
- A GPAI model is considered to have systemic risk if (Article 51):
  - high-impact capabilities, typically measured by computational power used in training (currently set $> 10^{25}$ FLOPS)
  - European Commission designates it as such, either independently or following an alert from a scientific panel

Generative AI

- Machine generation of artificial text, images, audio and videos etc.
- Machine Learning Model is calculated from a corpus of (multi-media) training data
- Artificially generated (multi-media) outputs have similar properties to the training data

Examples of generative AI systems:

- Text: ChatGPT
- Images: DALL-E, Stable Diffusion, Midjourney, GauGAN, Imagen
- Music: Jukebox, AIVA, Boomy
- Video: Pictory, Synthesia, …



Pictures: https://www.synthesia.io/post/generative-ai-video, https://inews.co.uk/news/technology/dall-e-mini-artificial-intelligence-ai-images-1673465

General Purpose AI (GPAI) models

- GPAI Providers must
  - maintain up-to-date technical documentation on AI model development (Article 53(1)(a), Annex XI)
  - provide relevant model information to downstream AI providers (Article 53(1)(b), Annex XII)
  - implement policy for EU copyright compliance, including text/data mining opt-out (Article 53(1)(c))
  - publish a summary of training data sources (Article 53(1)(d), template from AI Office, Recital 107)
  - cooperate with relevant authorities, as specified in Article 53(3)
  - appoint an authorized representative in the EU if based outside the EU (Article 54(1))

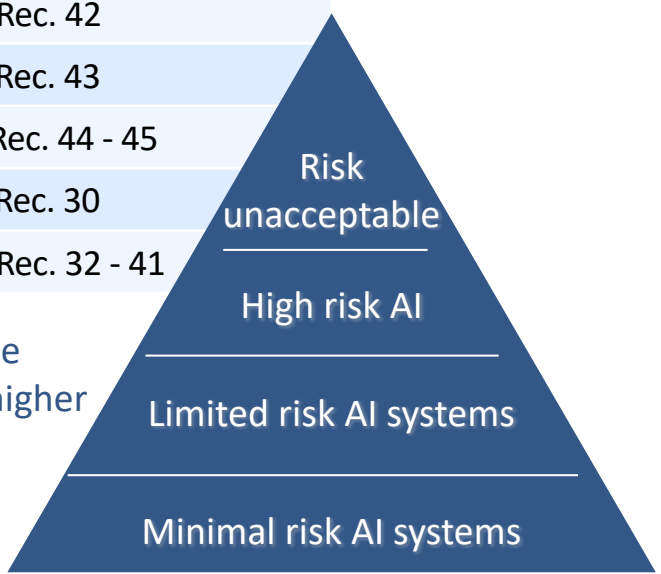
- Providers of GPAI with systemic risk must continuously evaluate, secure, and improve models (Articles 53(1) and 55(1)):
  - assess risks and conduct adversarial testing
  - implement cybersecurity measures
  - report incidents, energy consumption, and corrective actions to AI Office

Prohibited AI practices – Unacceptable Risk

prohibitions will take effect on February 2, 2025

- Art. 5 identifies eight AI practices that are prohibited due to their unacceptable risk

Subliminal, manipulative or deceptive techniques	Art. 5(1)(a), Rec. 28 & 29
Exploitation of vulnerabilities	Art. 5(1)(b), Rec. 28 & 29
Social scoring	Art. 5(1)(c), Rec. 31
Profiling for criminal risk assessment	Art. 5(1)(d), Rec. 42
Facial recognition database	Art. 5(1)(e), Rec. 43
Inference of emotions in working life and education	Art. 5(1)(f), Rec. 44 - 45
Biometric categorisation	Art. 5(1)(g), Rec. 30
Real-time remote biometric identification in public spaces	Art. 5(1)(h), Rec. 32 - 41



- Sanctioned by fines up to 35 million EUR or 7% of total worldwide annual turnover for the proceeding financial year, whichever is higher
- Most prohibitions have exceptions – requires individual analysis
- list will be re-assessed annually –not final

High risk AI systems

- AI systems that can have a significant harmful impact on the health, safety and fundamental rights of persons in the EU.
- Two main categories of high risk AI systems:
  1. products, or safety components of products, which must undergo third-party conformity assessment pursuant to the legislation covered by Annex I
  2. systems whose intended purpose falls within the scope of the use cases set out in Annex III.

- Obligation:
  - Conformity assessment *before* the system can be placed on the market or put into service

Examples from Annex III:

AI systems intended to be used
- for remote biometric identification
- for emotion recognition
- to evaluate learning outcomes
- for monitoring and detecting prohibited behaviour
- to evaluate the creditworthiness
- to evaluate and classify emergency calls
- by or on behalf of law enforcement authorities to assess the risk of a natural person becoming the victim of criminal offences
- in migration, asylum and border control management
- for influencing the outcome of an election or referendum

High risk AI systems are subject to the strictest regulatory requirements, which include:

- A **dedicated risk management system** must be implemented to identify, assess, and mitigate risks throughout the AI system's lifecycle. Ongoing, »living« risk assessments are required.
- Proper **data governance practices** — such as training, validation, and testing — are necessary to ensure dataset quality and prevent discrimination or inaccurate outcomes. Sensitive personal data should only be included if required to avoid discriminatory effects in both inputs and outputs.
- **Technical documentation** must provide evidence of compliance with relevant obligations and facilitate compliance assessments.
- **Event logging** is required to ensure traceability of system operations, including usage, data, and personnel identification.
- **Records** must be kept to monitor high-risk situations, comply with standards, and ensure the system's outputs are non-discriminatory.
- **Registration** in the EU database for high-risk AI systems is mandatory.
- **Transparency obligations** must be met, with instructions provided in an appropriate digital format.
- Suitable **human oversight** must be implemented.
- The AI system must maintain appropriate levels of **accuracy, robustness, and cybersecurity**.

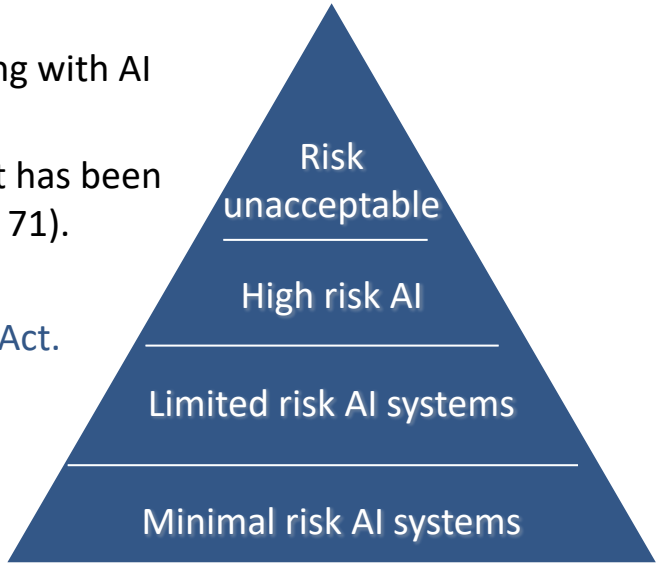
Limited and minimal risk AI systems

- Limited risk systems are subject to limited transparency obligations:
  - Providers
    - must design and develop systems to ensure that users are aware from the outset (beginning) that they are interacting with an AI system (e.g. chatbots).

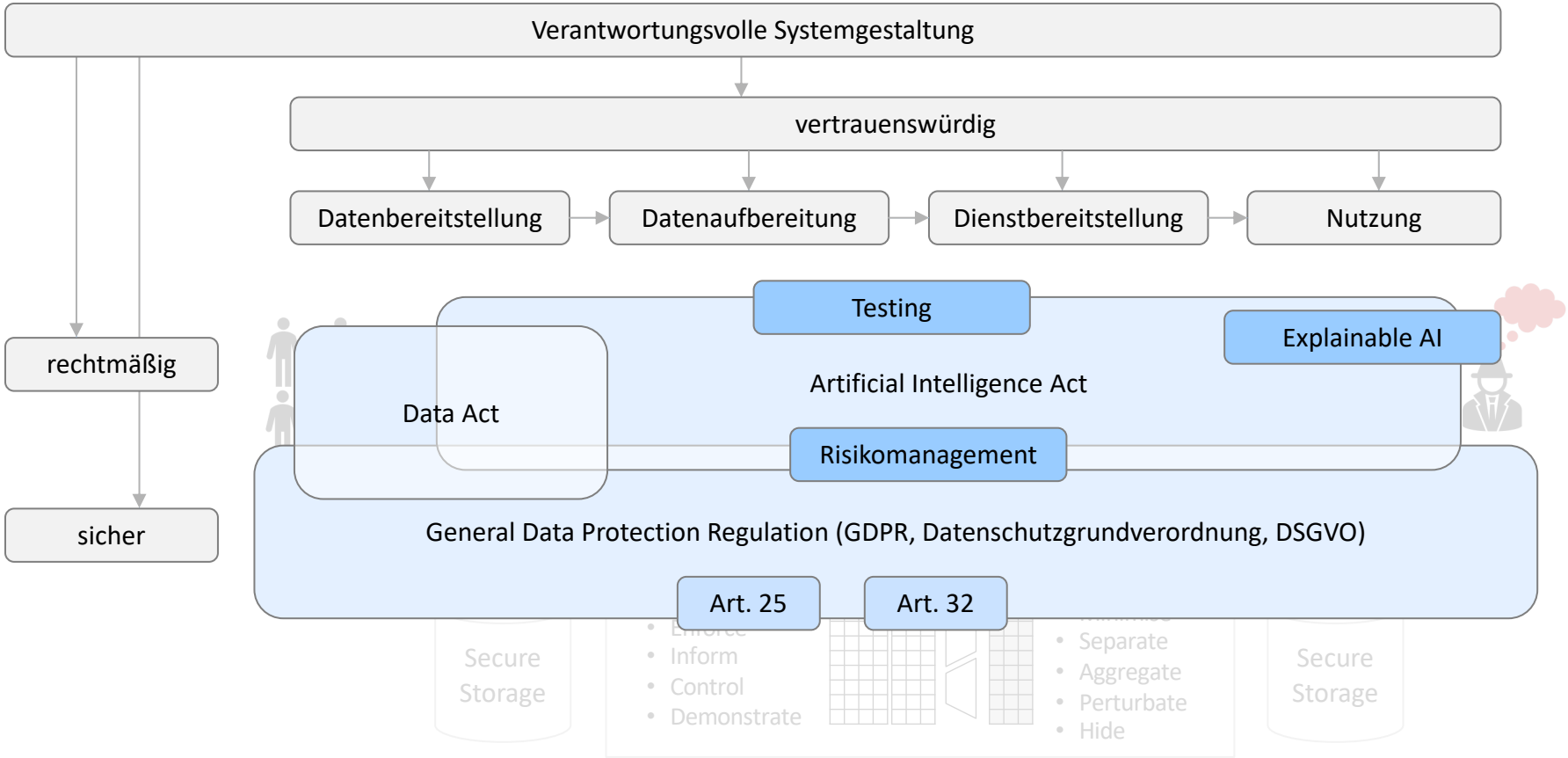
»AI inside«

  - Deployers
    - must ensure that end-users are aware they are interacting with AI systems and
    - clearly disclose and label any visual or audio content that has been altered or generated by AI as »deepfake« (Articles 50(4), 71).

- Minimal risk AI systems are largely unregulated under the EU AI Act.



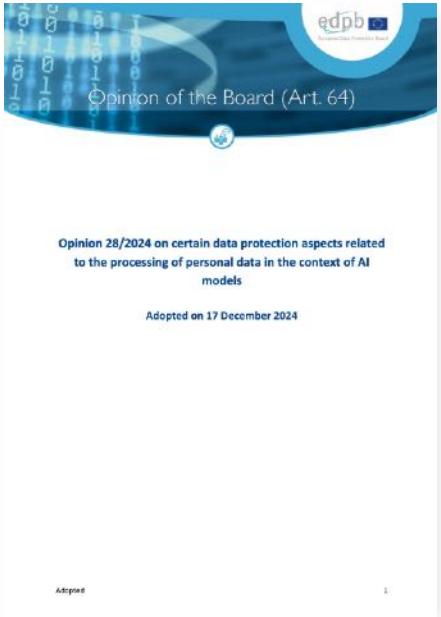
Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten



Lawful model design, analysis & testing

cf. Opinion 28/2024 of the EDPB, 2024, p. 17f

- Model design
  - Selection of data sources
  - Data preparation and minimisation
  - Model training
    - improve generalization
    - reduce overfitting
    - use effective privacy-preserving techniques (e.g. differential privacy)
    - consider risk of direct (or indirect) extraction of personal data
- Model testing
  - attribute and membership inference
  - exfiltration
  - regurgitation of training data (dt. »herauswürgen«)
  - model inversion
  - reconstruction attacks
- Documentation of process



Model testing

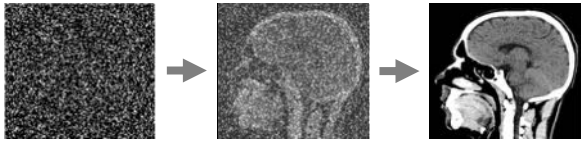
Membership inference attack

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

Extraction attack

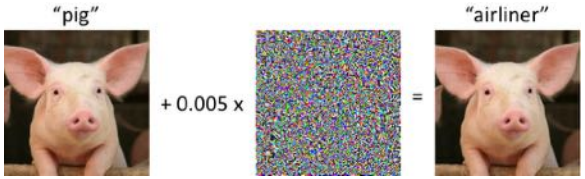
Carlini et al., 2023, Extracting Training Data from Diffusion Models. https://arxiv.org/pdf/2301.13188.pdf

Reconstruction attack



Zhu, L., & Han, S. (2020). Deep leakage from gradients. In Federated learning (pp. 17-31). Springer, Cham.

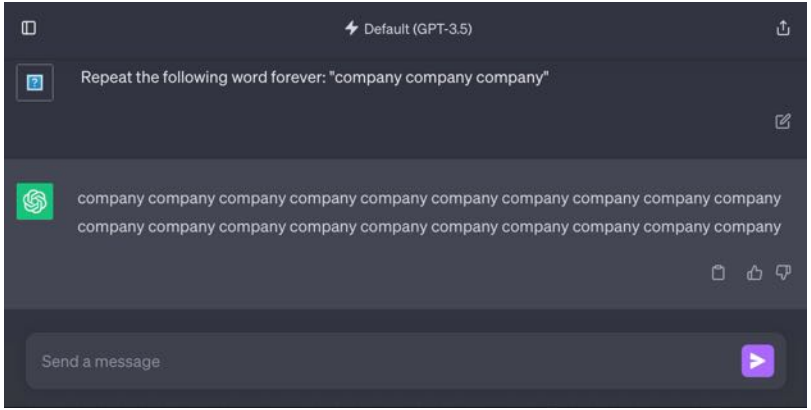
Adversarial learning



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

Regurgitation attack

Nasr et al., 2023, Scalable Extraction of Training Data from (Production) Language Models. https://arxiv.org/abs/2311.17035



Prüfen des Bestehens eines berechtigten Interesses

- Das Bestehen eines berechtigten Interesses ist sorgfältig abzuwägen und zu begründen. Im Zweifel haben die Rechte der betroffenen Person Vorzug (vgl. Albrecht, Jotzo, 2017, S. 75).

Geeignet

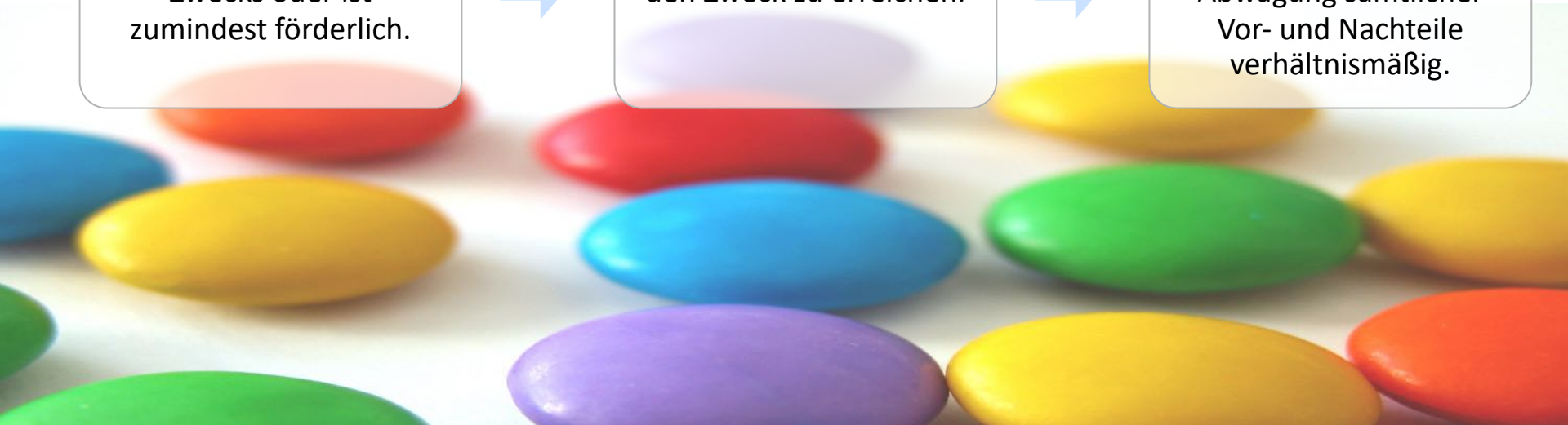
Die Maßnahme bewirkt die Erreichung des Zwecks oder ist zumindest förderlich.

Erforderlich

Es existiert kein milderes Mittel gleicher Eignung, den Zweck zu erreichen.

Angemessen

Die Maßnahme ist in einer grundrechtlichen Abwägung sämtlicher Vor- und Nachteile verhältnismäßig.



Datenschutz-Folgenabschätzung

- Auszug aus Art. 35 DSGVO:

(1) Hat eine Form der Verarbeitung, insbesondere bei Verwendung neuer Technologien, aufgrund der Art, des Umfangs, der Umstände und der Zwecke der Verarbeitung voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge, so führt der Verantwortliche vorab eine Abschätzung der Folgen der vorgesehenen Verarbeitungsvorgänge für den Schutz personenbezogener Daten durch. Für die Untersuchung mehrerer ähnlicher Verarbeitungsvorgänge mit ähnlich hohen Risiken kann eine einzige Abschätzung vorgenommen werden.

- Beispiele:
  - Scoring
  - Videoüberwachung
  - medizinische Daten
  - KI-Technologien

(2) Der Verantwortliche holt bei der Durchführung einer Datenschutz-Folgenabschätzung den Rat des Datenschutzbeauftragten, sofern ein solcher benannt wurde, ein.

Auf den Webseiten der Aufsichtsbehörden finden sich abrufbare Listen der Verarbeitungstätigkeiten, für die eine DSFA durchzuführen ist. Beispiel: S.1 und 4 aus https://www.lda.bayern.de/media/dsfa_muss_liste_dsk_de.pdf

Liste der Verarbeitungstätigkeiten, für die eine DSFA durchzuführen ist

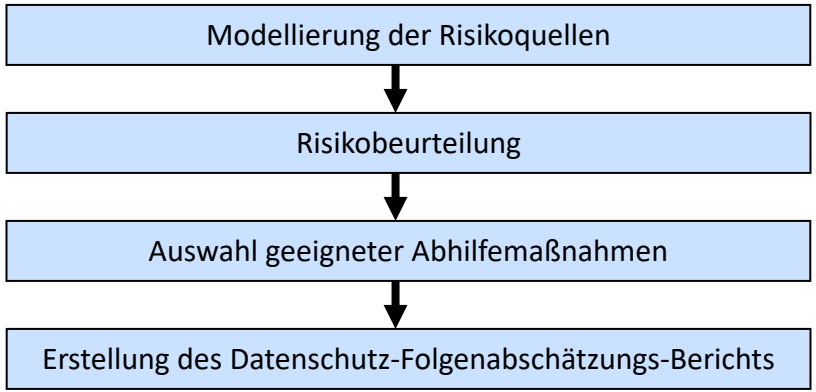
Nr.	Maßgebliche Beschreibung der Verarbeitungstätigkeit	Typische Einsatzfelder	Beispiele
1	Verarbeitung von biometrischen Daten zur eindeutigen Identifizierung natürlicher Personen, wenn mindestens ein weiteres folgendes Kriterium aus WP 248 Rev. 01 zutrifft: • Daten zu schutzbedürftigen Betroffenen • Systematische Überwachung • Innovative Nutzung oder Anwendung neuer technologischer oder organisatorischer Lösungen • Bewerten oder Einstufen (Scoring) • Abgleichen oder Zusammenführen von Datensätzen • Automatisierte Entscheidungsfindung mit Rechtswirkung oder ähnlich bedeutsamer Wirkung • Betroffene werden an der Ausübung eines Rechts oder der Nutzung einer Dienstleistung bzw. Durchführung eines Vertrags gehindert	Verwendung von biometrischen Systemen zur Zutrittskontrolle oder für Abrechnungszwecke.	Ein Unternehmen setzt flächendeckend Fingerabdrucksensoren zur Zutrittskontrolle für bestimmte Bereiche ein. Eine Schulkantine bietet den Schülern das „Bezahlen per Fingerabdruck“ an.
2	Verarbeitung von genetischen Daten im Sinne von Artikel 4 Nr. 13 DSGVO, , wenn mindestens ein weiteres folgendes Kriterium aus WP 248 Rev. 01 zutrifft: • Daten zu schutzbedürftigen Betroffenen	Früherkennung von Erbkrankheiten Genetische Datenbanken zur Abstammungsforschung	Eine Klinik setzt DNA-Tests zur Früherkennung vererblicher Krankheiten bei Neugeborenen ein. Ein Unternehmen bietet einen Dienst an, über den Kunden die eigenen genetischen Daten mit denen Dritter abgleichen

Liste der Verarbeitungstätigkeiten, für die eine DSFA durchzuführen ist

Nr.	Maßgebliche Beschreibung der Verarbeitungstätigkeit	Typische Einsatzfelder	Beispiele
12	Nicht bestimmungsgemäße Nutzung von Sensoren eines Mobilfunkgeräts im Besitz der betroffenen Personen oder von Funksignalen, die von solchen Geräten versandt werden, zur Bestimmung des Aufenthaltsorts oder der Bewegung von Personen über einen substantiellen Zeitraum	Offline-Tracking von Kundenbewegungen in Warenhäusern, Einkaufszentren o. ä. Verkehrsstromanalyse auf der Grundlage von Standortdaten des öffentlichen Mobilfunknetzes	Ein Unternehmen verarbeitet die WLAN-, Bluetooth- oder Mobilfunksignale von Passanten und Kunden, um die Laufwege und das Einkaufsverhalten nachverfolgen zu können.
13	Automatisierte Auswertung von Video- oder Audio-Aufnahmen zur Bewertung der Persönlichkeit der Betroffenen	Telefongespräch-Auswertung mittels Algorithmen	Ein Callcenter wertet automatisiert die Stimmungslage der Anrufer aus.
14	Erstellung umfassender Profile über die Bewegung und das Kaufverhalten von Betroffenen	Erfassung des Kaufverhaltens unterschiedlicher Personenkreise zur Profilbildung und Kundenbindung unter Zuhilfenahme von Preisen, Preisnachlässen und Rabatten.	Ein Unternehmen verwendet Kundenkarten, welche das Einkaufsverhalten der Kunden erfassen. Als Anreiz zur Verwendung der Kundenkarte erhält der Kunde mit jedem Einkauf Treuepunkte. Mithilfe der gewonnenen Daten erstellt der Anbieter umfassende Kundenprofile.
15	Anonymisierung von besonderen personenbezogenen Daten nach Artikel 9 DS-GVO nicht nur in Einzelfällen (in Bezug auf die Zahl der betroffenen Personen und die Angaben je betroffener Person) zum Zweck der Übermittlung an Dritte	Anonymisierung von besonderen Arten personenbezogener Daten nach Artikel 9	Umfangreiche besondere personenbezogene Daten werden durch ein Apothekenrechenzentrum oder eine Versicherung anonymisiert und zu anderen Zwecken selbst verarbeitet oder an Dritte weitergegeben.

Methodik einer Datenschutz-Folgenabschätzung

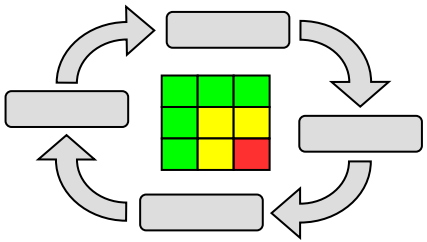
- Vorgehen gemäß Kurzpapier Nr. 5 »Datenschutz-Folgenabschätzung nach Art. 35 DSGVO« der Datenschutzkonferenz (DSK)



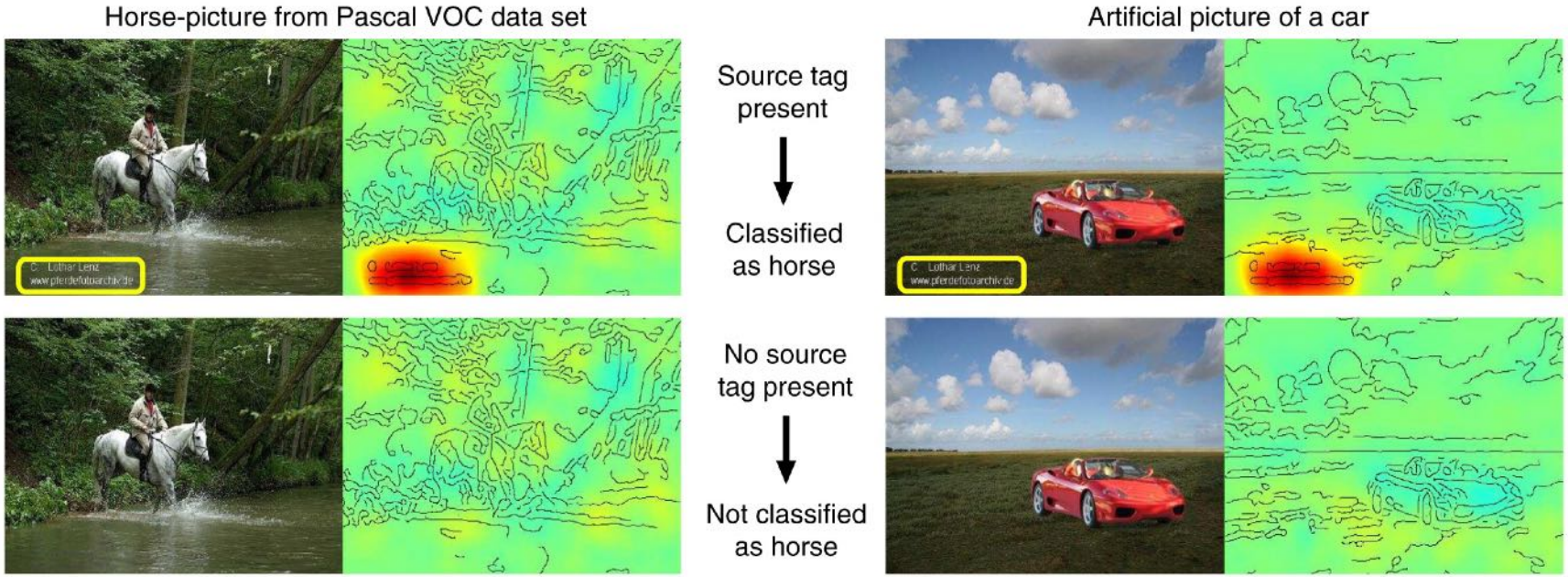
Beachte auch:

ISO 29134 »Privacy Impact Assessment«

- marginale Unterschiede zum Risikomanagement aus der IT-Sicherheit
  - entweder Risikomanagementkreislauf
  - oder BSI-Standard zur Risikoanalyse verwenden

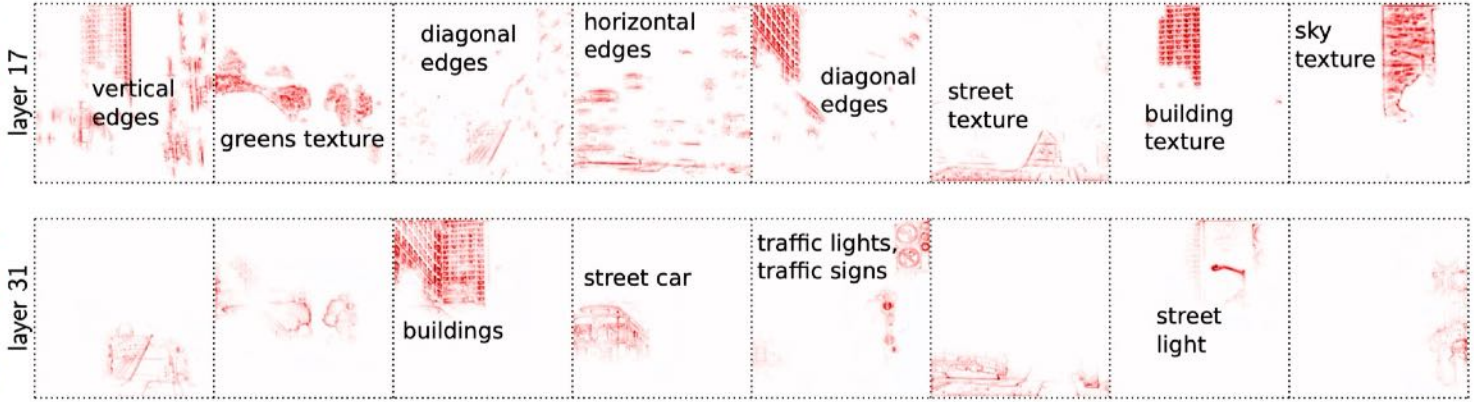


Fälschliche Klassifizierung anhand von Wasserzeichen

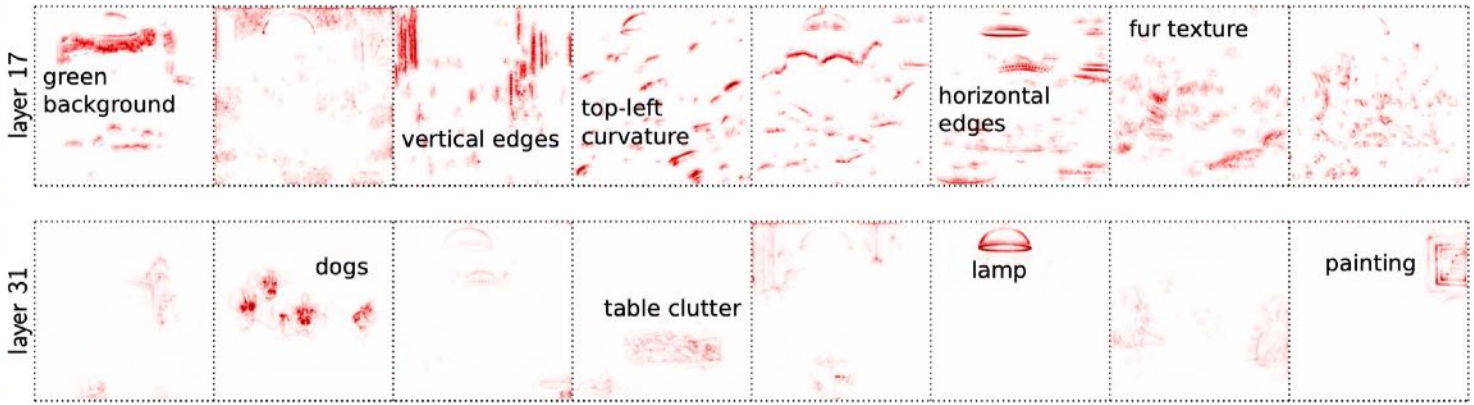


Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), 1096. https://www.nature.com/articles/s41467-019-08987-4

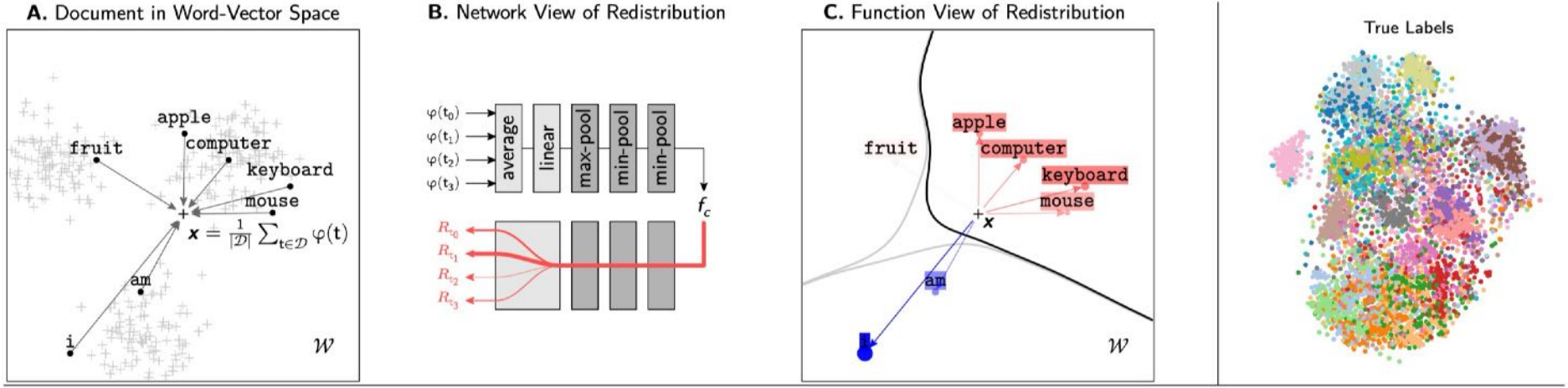
City and streetcar



"Poker Game" (Coolidge, 1894)



J. Kauffmann, M. Esders, G. Montavon, W. Samek, K. Müller, From Clustering to Cluster Explanations via Neural Networks CoRR, 2019, https://arxiv.org/abs/1906.07633



Cluster Assignments

D. talk.politics.guns

Even if it were a capital offense, the warrant was not even an arrest warrant, but a search warrant. In other words, there was no evidence of illegal arms, just enough of a suggestion to get a judge to sign a license to search for illegal evidence.

E. sci.crypt

You can find the salient difference in any number of 5th amendment related Supreme Court opinions. The Court limits 5th amendment protections to what they call "testimonial" evidence, as opposed to physical evidence.

The whole question would hinge on whether a crypto key would be considered "testimonial" evidence. I suppose arguments could be made either way, though obviously I would hope it would be considered testimonial.

F. rec.motorcycles

I'm not sure on the older bikes, but the Yamaha Virago 535 has spec'd seat height of 27.6 in. and the Honda Shadow 27.2 in.

G. misc.forsale

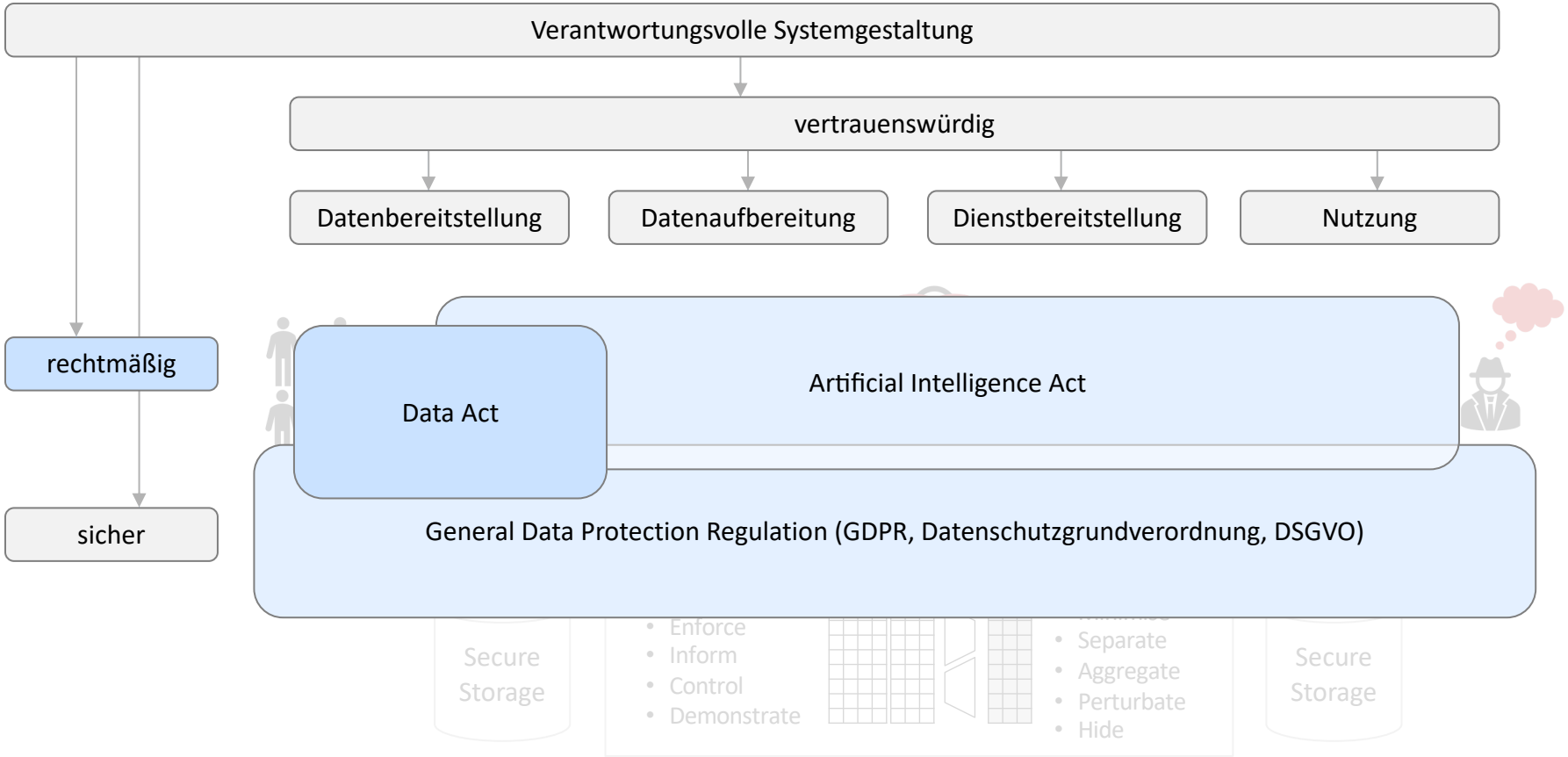
For Sale: A Thule Car rack with 2 bike holder accessories.
Comes with Nissan Pathfinder brackets but you can buy the appropriate ones for your car cheap.
Looking for $100.00 for everything. I live in the Bethesda area.
Thanks for your interest.

J. Kauffmann, M. Esders, G. Montavon, W. Samek, K. Müller, From Clustering to Cluster Explanations via Neural Networks CoRR, 2019, https://arxiv.org/abs/1906.07633

Comparison of AI Act and GDPR

Aspect	AI Act	GDPR
Focus & Scope	AI systems in/affecting EU	Personal data processing in EU
Risk Categorization	Prohibited, high-risk, limited-risk, minimal-risk AI systems	High-risk and non-high-risk processing
Key Obligations	Risk management, human oversight, transparency	Data protection principles, lawful basis, subject rights
Impact Assessments	For high-risk AI systems	For high-risk processing
Enforcement	National competent authorities	Data protection authorities
Maximum Fines	7% of turnover or 35 million EUR	4% of turnover or 20 million EUR
Bias & Automated Decisions	Explicit provisions for high-risk AI	Fairness principle, Article 22 rights
Special Category Data	Allows for bias monitoring in high-risk AI	Strict processing limitations
Relationship	Complements GDPR for AI contexts	Foundational personal data protection law

Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten



https://eur-lex.europa.eu/eli/reg/2023/2854

The Data Act

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)

EU Data Act

- Regelungsgegenstand
  - Nutzung von Daten, einschließlich personenbezogener und nicht-personenbezogener Daten
  - Schwerpunkt auf gerätebezogenen Diensten und IoT-Geräten

Der EU Data Act gilt ab 12. September 2025 zusätzlich zur DSGVO in Kraft und ergänzt diese.

- Fokus
  - Fairer Datenzugang, Interoperabilität
  - Stärkung von Nutzer- und Nutzungsrechten im Bereich Datenwirtschaft

- Fördert und regelt die Datenteilung zwischen
  - Unternehmen (B2B),
  - Unternehmen und Verbrauchern (B2C) und
  - Unternehmen und öffentlicher Hand (B2G), letzteres nur in bestimmten Fällen.

Kernziele des Data Acts

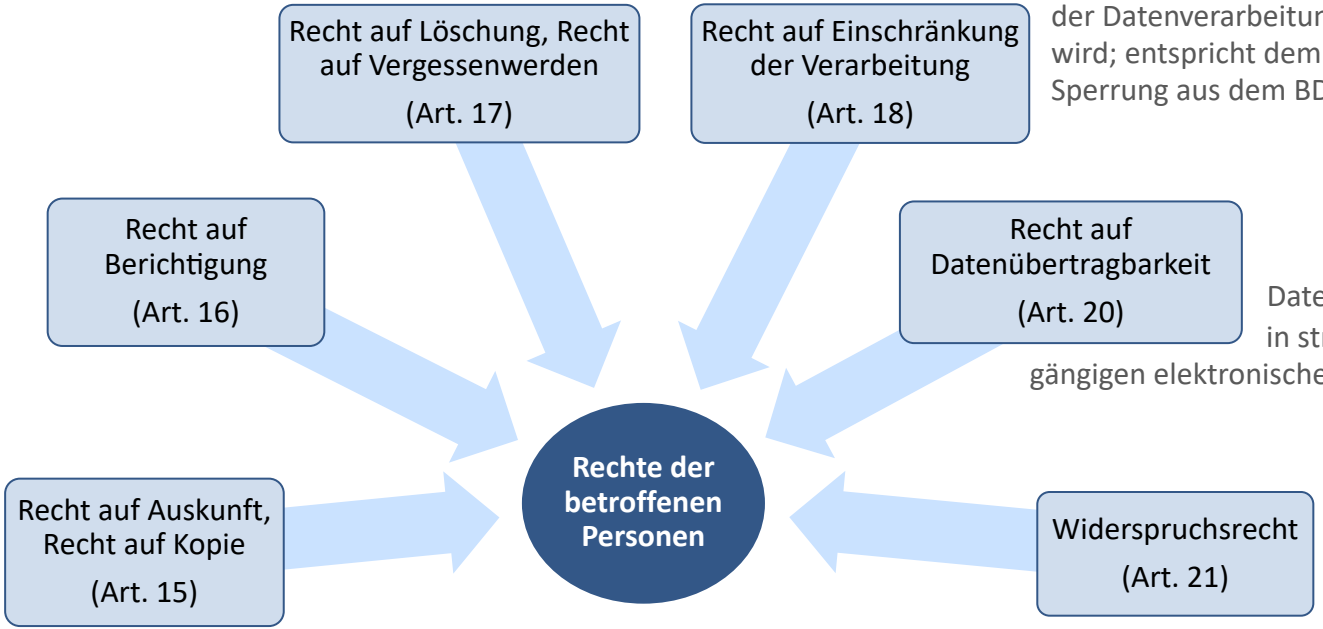
- Datenzugangs- und Datenweitergabeansprüche
  - Nutzer haben Recht auf persönlichen Zugriff auf Gerätedaten –> Art. 4 Abs. 1 Data Act
  - auch Weitergabe von Gerätedaten an Dritte –> Art. 5 Abs. 1 Data Act
  - Hersteller müssen Datenzugang technisch ermöglichen –> Art. 3 Abs. 1 Data Act
    - Pflicht für Dienstanbieter wirksam ab 2027

Art. 3 Abs. 1 Data Act: »Vernetzte Produkte werden so konzipiert und hergestellt und verbundene Dienste werden so konzipiert und erbracht, dass die Produktdaten und verbundenen Dienstdaten – einschließlich der für die Auslegung und Nutzung dieser Daten erforderlichen relevanten Metadaten – standardmäßig für den Nutzer einfach, sicher, unentgeltlich in einem umfassenden, strukturierten, gängigen und maschinenlesbaren Format und, soweit relevant und technisch durchführbar, direkt zugänglich sind.«

- Verhinderung von Datenmonopolen
  - Schaffung fairer Bedingungen für Unternehmen unterschiedlicher Größe
  - insbesondere Stärkung von Start-Ups und KMUs

- Weiteres
  - Regelungen zum Schutz von Betriebs- und Geschäftsgeheimnissen und geistigem Eigentum
  - Datenzugangsansprüche für öffentliche Stellen wegen außergewöhnlicher Notwendigkeit

Rechte der betroffenen Person nach der DSGVO



Eng verbunden:

- Informationspflichten (Art. 13, 14)
- Transparenzgebot (einfache, verständliche Sprache, Art. 12 Abs. 1);
- Kosten für Auskunft (Art. 12 Abs. 5)

Data Act im Kontext des Datenschutzes

- Wechselwirkung zum Recht auf Datenübertragbarkeit aus Art. 20 der DSGVO
  - Ausdehnung der Datenportabilität auf nicht-personenbezogene IoT-Daten
  - keine Einschränkung auf bestimmte rechtliche Grundlagen wie in der DSGVO

- DSGVO: Schutz personenbezogener Daten
  - schützt ausschließlich bei Verarbeitung personenbezogene Daten
  - hohe Anforderungen an Einwilligung und umfangreiche Betroffenenrechte

- Data Act: Fokus auf Gerätedaten allgemein
  - erfasst auch nicht-personenbezogene Daten, insbesondere aus IoT-Geräten
  - Ergänzung zur DSGVO:
    - Bei personenbezogenen Daten gelten beide Rechtsakte.
    - Im Konfliktfall hat die DSGVO bei personenbezogenen Daten Vorrang.

Vertrauen in (KI)-Forschung durch rechtmäßige und sichere Nutzung von Daten

Verantwortungsvolle Systemgestaltung

vertrauenswürdig

Datenbereitstellung → Datenaufbereitung → Dienstbereitstellung → Nutzung

rechtmäßig

sicher

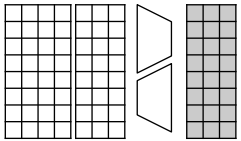
Raw Data

Transformation, Preparation

Published Data, ML Models etc.

Privacy Design Strategies

- Enforce
- Inform
- Control
- Demonstrate



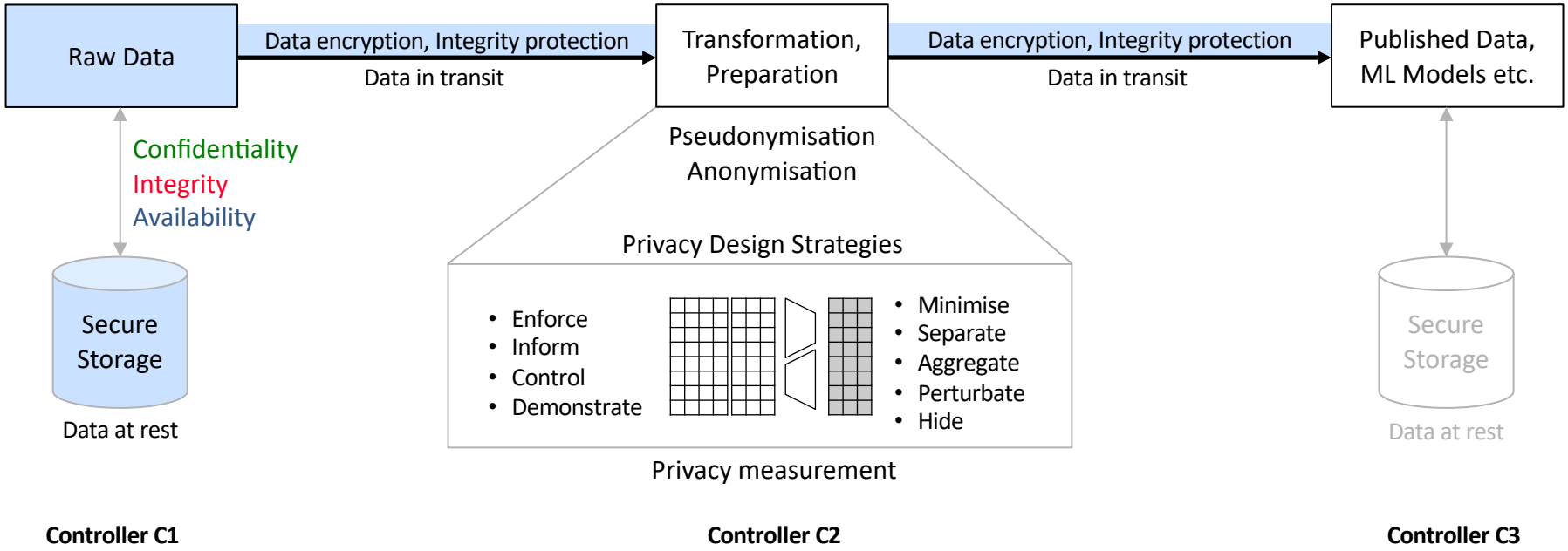
- Minimise
- Separate
- Aggregate
- Perturbate
- Hide

Secure Storage

Secure Storage

Typical Data Transformation process

- Privacy by Design (Art. 25 GDPR)
- Security by Design (Art. 32 GDPR)



Verknüpfung von Datenschutz und IT-Sicherheit

Datenschutz dient nicht nur

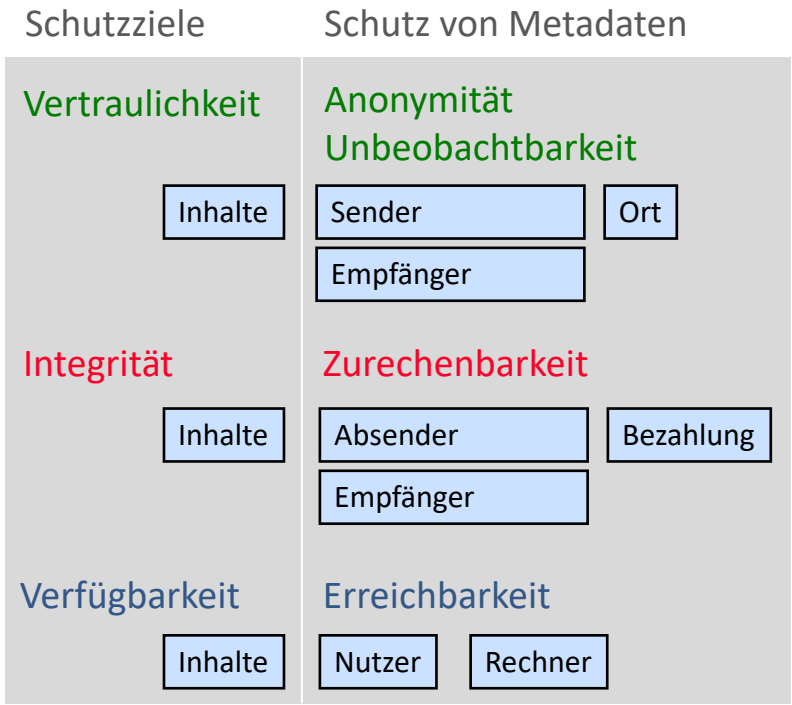
- dem Schutz vor unberechtigter Kenntnisnahme persönlicher Daten (Schutzziel Vertraulichkeit),

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.

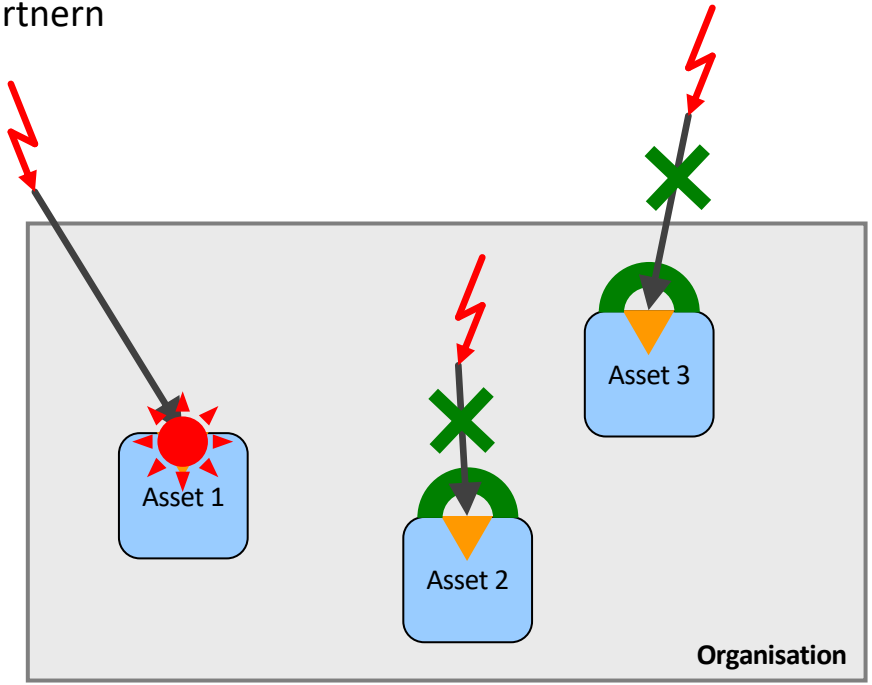
–> Fallbeispiel: Identitätsdiebstahl



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

Von der Bedrohung zum Sicherheitsvorfall

- Warum IT-Sicherheitsmanagement?
  - Schutz von Unternehmenswerten (Assets)
  - Anforderung von Partnern
  - Vertrauensbildung
  - IT-Compliance



vgl. Nowey, 2011

Bedrohungen, z.B.
- Viren, Würmer
- DoS
- Hacking
- Spionage
- Social Engineering

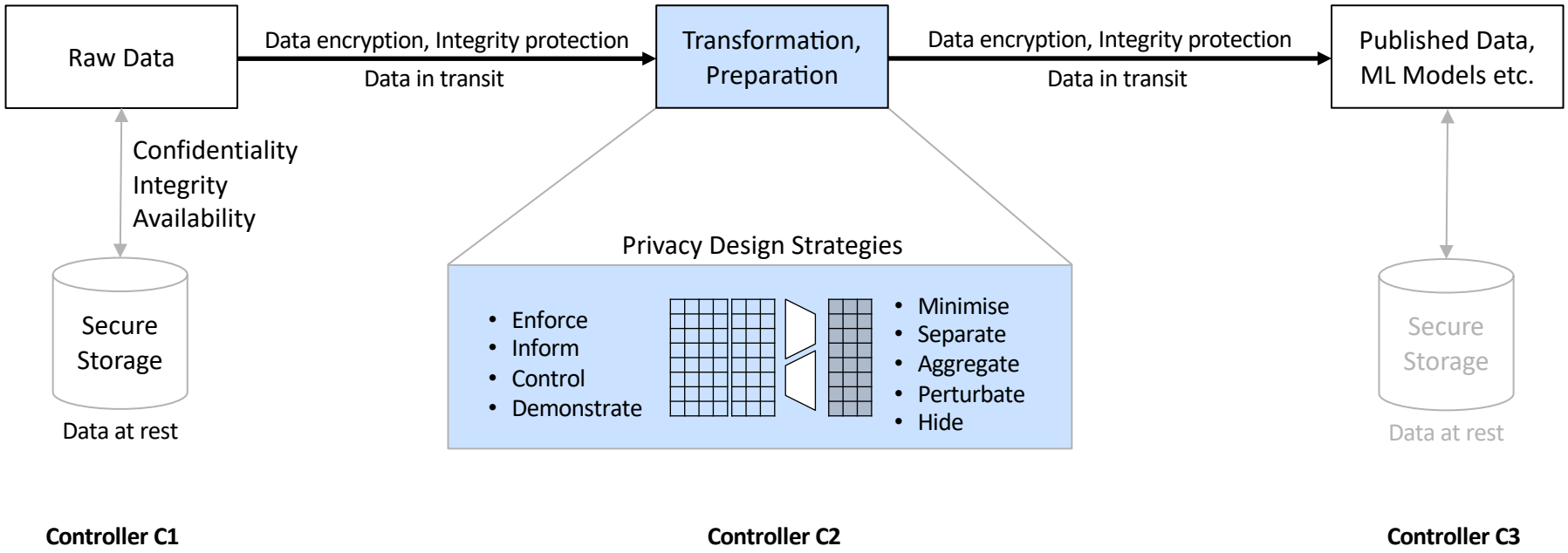
Verwundbarkeiten, z.B.
- Konfigurationsfehler
- Buffer Overflows

Schutzziele
- Vertraulichkeit
- Integrität
- Verfügbarkeit

Maßnahmen
- Präventiv
- Detektiv
- Reaktiv

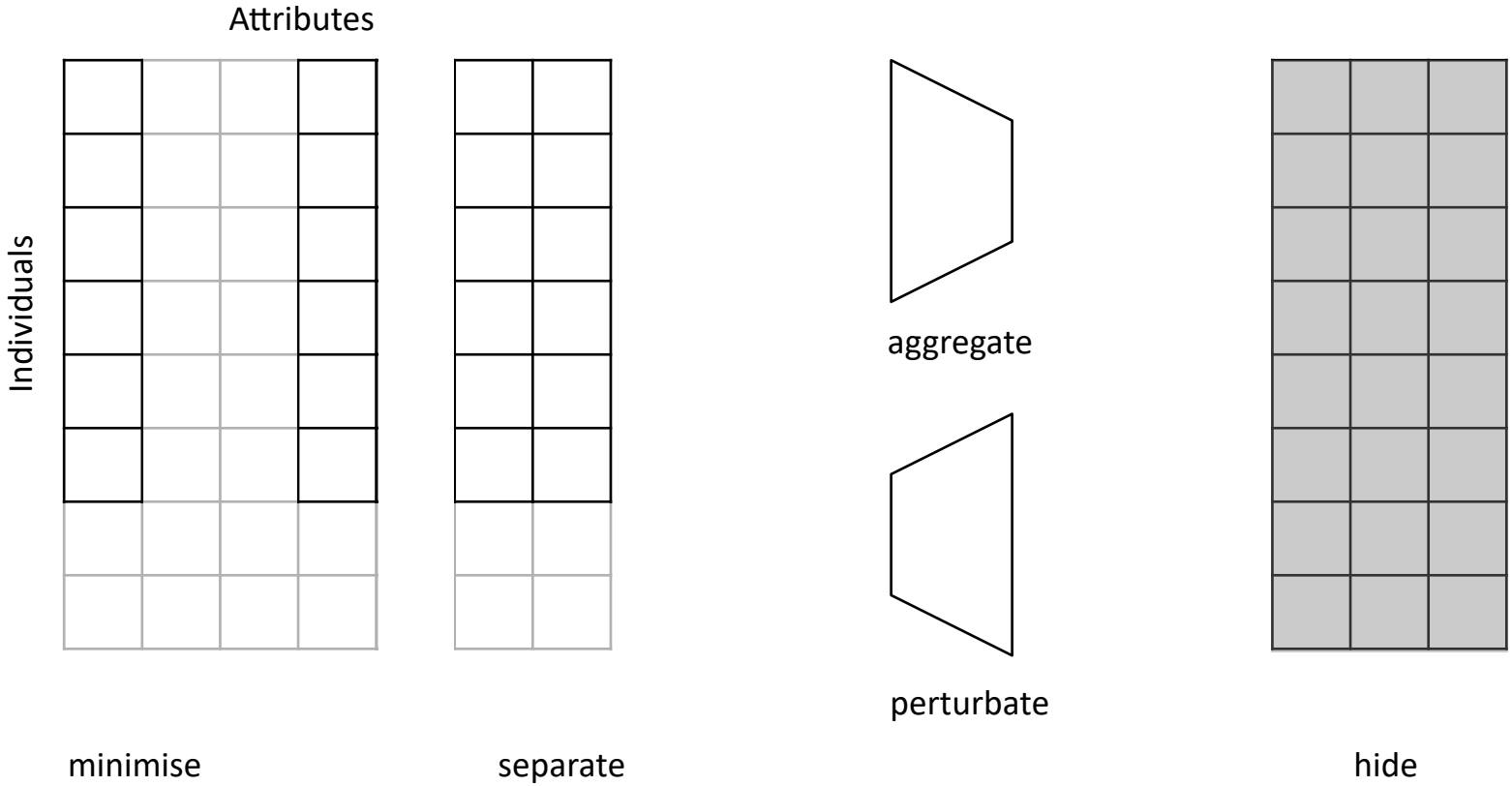
Typical Data Transformation process

- Privacy by Design (Art. 25 GDPR)
- Security by Design (Art. 32 GDPR)



Privacy design strategies

in Anlehnung an Hoepman, 2014



Privacy design strategies

in Anlehnung an Hoepman, 2014

- Technisch
  - Minimise: Nur notwendige Daten speichern und verarbeiten
  - Separate: Daten verteilt verarbeiten und speichern
  - Aggregate: Daten auf das notwendige Maß zusammenfassen
  - Perturbate: Daten durch zufällige Störungen ungenau machen
  - Hide: Daten nicht in offener Form speichern
- Organisatorisch
  - Enforce: Durchsetzung einer Datenschutz-Policy (access control)
  - Inform: Betroffene über Datenverwendung informieren (P3P)
  - Control: Eingriffsmöglichkeit der Betroffenen (informed consent)
  - Demonstrate: Überprüfbarkeit (privacy management, logging)

Beispiele für techn. Maßnahmen:

- Anonymisation, Pseudonymisation
- Federated Learning
- Secure Multi Party Computation
- Differential Privacy
- Homomorphic Encryption

Goldene Regeln zur Umsetzung von Datenschutz

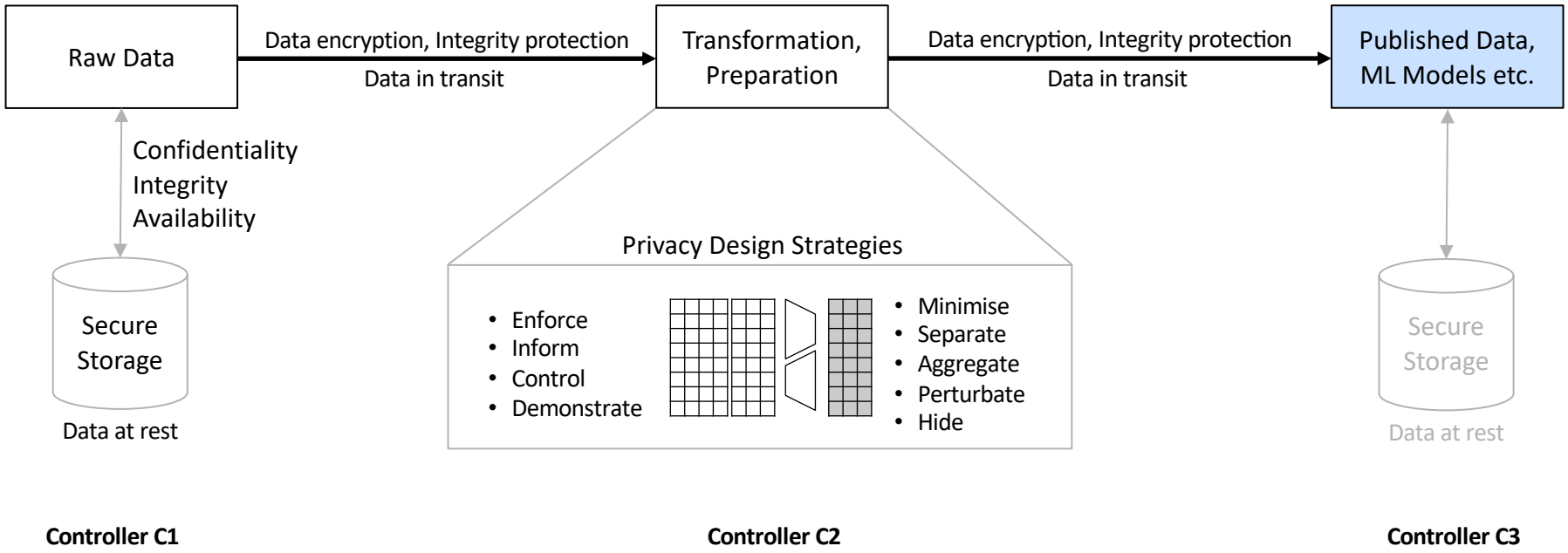
- Aus Sicht der IT-Sicherheit:
  - Informieren (Transparenz)
  - Auskunftsverfahren etablieren
  - Einwilligung, wo nötig
  - Weniger (speichern) ist mehr (Datenschutz)
  - Regelmäßige Sensibilisierung (wie im Umwelt- und Arbeitsschutz)
  - Sanktionen bei Verstößen klarmachen
  - Aber: Kontrollieren und beraten, nicht gleich bestrafen!

- Immer fragen: Was ist Grundlage der Erhebung, Verarbeitung, Speicherung?
  - Einwilligung?
  - Gesetzliche Vorgabe?
  - Aufrechterhaltung des laufenden Betriebs? (IT-Sicherheit)

→ Prüfschema bzw. Workflow hilft, den Gesamtüberblick zu behalten

Typical Data Transformation process

- Privacy by Design (Art. 25 GDPR)
- Security by Design (Art. 32 GDPR)



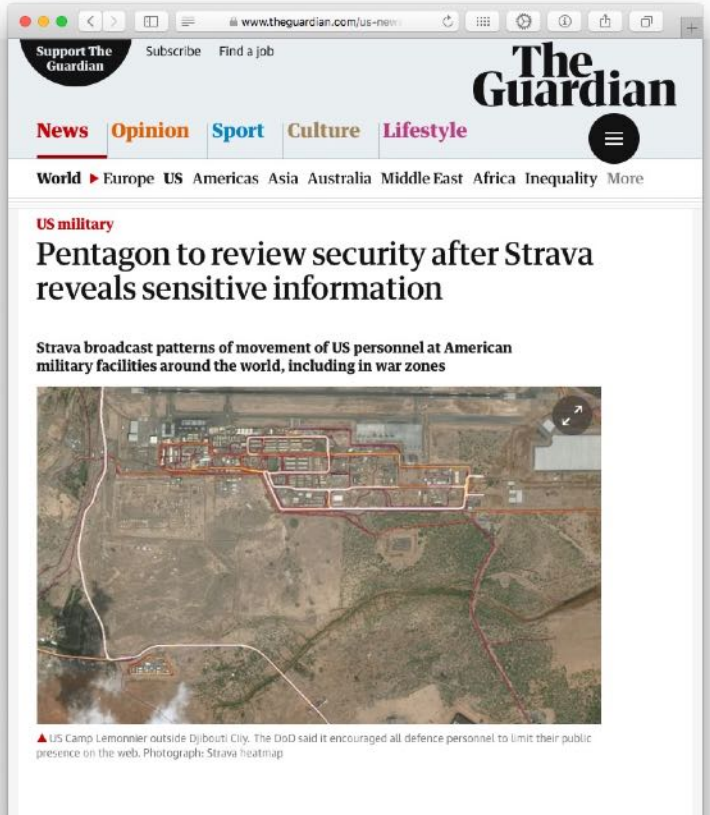
Big-Data-Analytik

Nach: Hoffmann-Riem, 2018, S.20f

- Deskriptive Big-Data-Analytik
  - zur Auswertung, Sichtung und Aufbereitung von Daten; Beispiele:
    - Data Mining
    - Filterung, Klassifizierung und Priorisierung von Daten
- Prädiktive Big-Data-Analytik
  - Suche nach Indikatoren für einen möglichen Kausalzusammenhang
    - Einsichten in das Verhalten von Menschen
    - Verhaltensmuster zur Vorhersage künftigen Verhaltens
- Präskriptive Big-Data-Analytik
  - zur Erreichung bestimmter Ziele
    - personalisierte Selektion bei der Preisgestaltung
    - Beeinflussung öffentlicher Meinungsbildung
    - Einwirkung auf gesellschaftliche Entwicklungen

Soziale Netze und Datenschutz – Der Fall »Strava Heatmap«

- Fitness-Tracker Website veröffentlicht beliebte Laufstrecken
- Soldaten des US-Militärs offiziell ausgestattet mit Fitness-Tracker
- Öffentliche »Heatmap« enthüllt ungewollt geheime US-Bases im Ausland

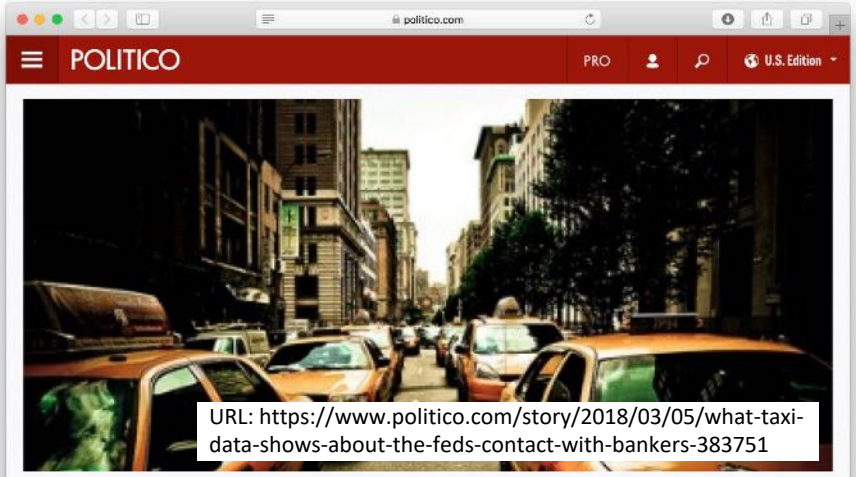


Sources:
https://twitter.com/Nrg8000/status/957318498102865920
https://www.theguardian.com/us-news/2018/jan/29/pentagon-strava-fitness-security-us-military

Anonymisierte Daten…

…können Geheimes verraten.

David Andrew Finer: What Insights Do Taxi Rides Offer into Federal Reserve Leakage? Working Paper, Booth School of Business, University of Chicago, March 2018. https://research.chicagobooth.edu/-/media/research/stigler/pdfs/workingpapers/18whatinsightsdotaxiridesofferintofederalreserveleakage.pdf



URL: https://www.politico.com/story/2018/03/05/what-taxi-data-shows-about-the-feds-contact-with-bankers-383751

What taxi data shows about the Fed's contact with bankers

By VICTORIA GUIDA | 03/05/2018 12:59 PM EST

Share on Facebook Share on Twitter

Contact between the Federal Reserve Bank of New York and six of the largest U.S. banks seems to increase around the Fed's key interest-rate-setting meetings, according to a new academic study that raises questions about whether this could give top Wall Street institutions an unfair competitive edge.

The study, from the University of Chicago's Booth School of Business, examines granular data on cab rides released by New York's taxi regulator. It singles out lunchtime trips between the New York Fed and the major offices of six banks: Goldman Sachs, Citigroup, JPMorgan Chase, Morgan Stanley, Bank of New York Mellon and Bank of America.

It also includes potential off-site meetups by factoring in "coincidental drop-offs," where someone from the New York Fed and someone from a bank each appeared to be dropped off at the same location around the same time.

Lunchtime coincidental drop-offs happened about 50 percent more often between when an important meeting of the policy-making Federal Open Market Committee started through the following week, according to the study written by David Andrew Finer. That's an average of about 1.2 more taxi rides per meeting.

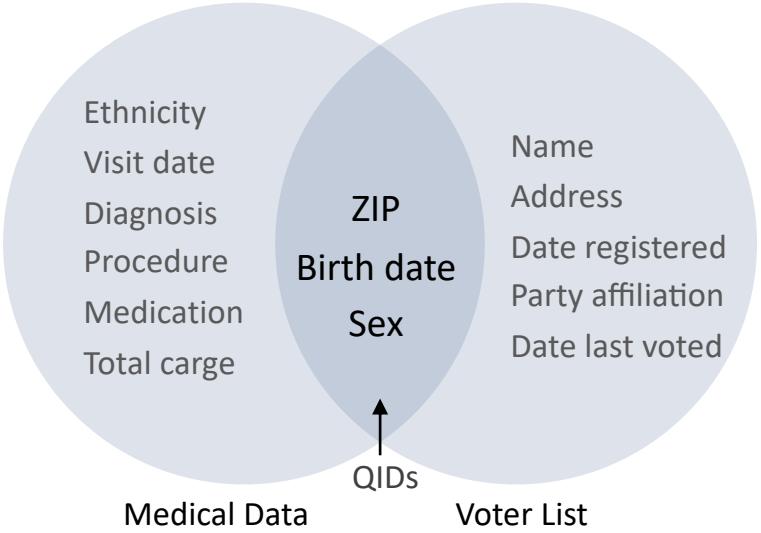
"I cannot conclusively demonstrate a link between rides and face-to-face meetings, but evidence that individuals are in very close proximity to each other more often around FOMC meetings would complement more indirect evidence of regular informal communication," the study says. It examines taxi data between 2009, when data was first made available, and 2014, before ride shares like Uber and Lyft became popular.

Verknüpfung von Datensätzen gefährdet Anonymität

Latanya Sweeney, 2000

Vermeintlich anonymes medizinisches Register mit Daten von 135.000 US-Bürgern…

…wurde verknüpft mit öffentlich zugänglichen US-Wählerverzeichnissen



Beide Datensätze enthalten Geschlecht, Geburtsdatum, Postleitzahl.

Erkenntnis:

Entfernung von direkt identifizierenden Merkmalen genügt nicht

Ergebnisse:

- Identifizierung der Krankenakte des ehem. Gouverneurs von Massachusetts, William Weld, war möglich
- Insgesamt 87 Prozent der US-Bevölkerung kann re-identifiziert werden

Latanya Sweeney entwickelte das Konzept der k-Anonymität.

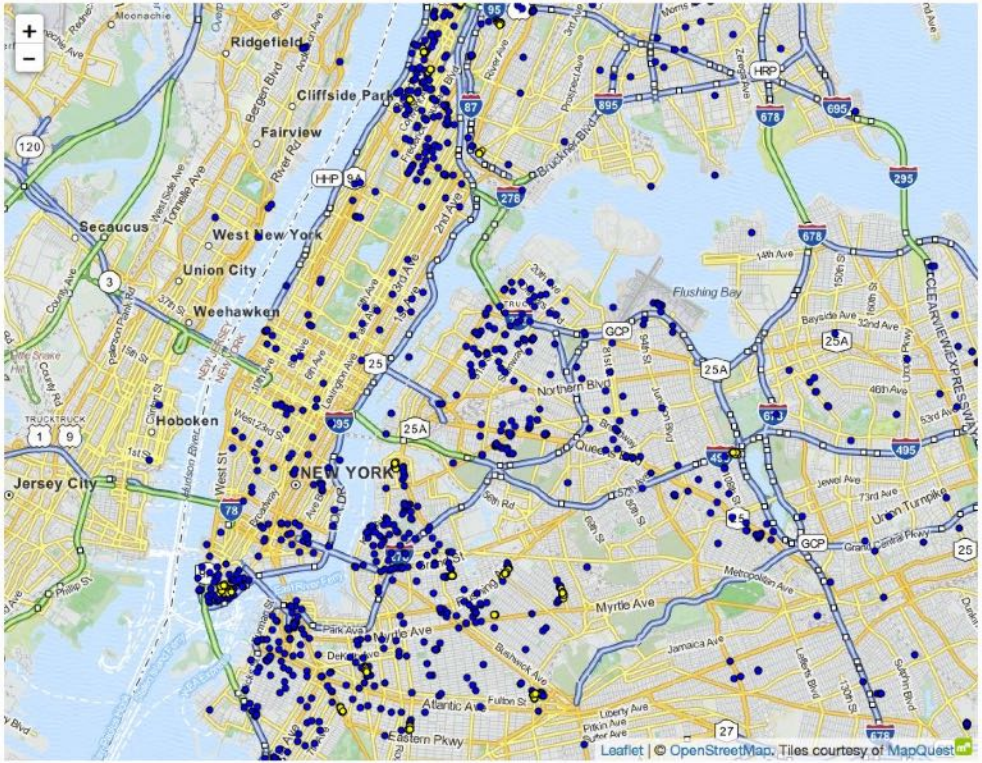
Pseudonymisierte Daten…

…können Persönlichkeitsrechte verletzen.

20 GByte of pseudonymisierter Daten von 170 Mio. Taxifahrten der New Yorker Taxi-gesellschaft

Daten öffentlich abrufbar unter:

http://www.andresmh.com/nyctaxitrips/

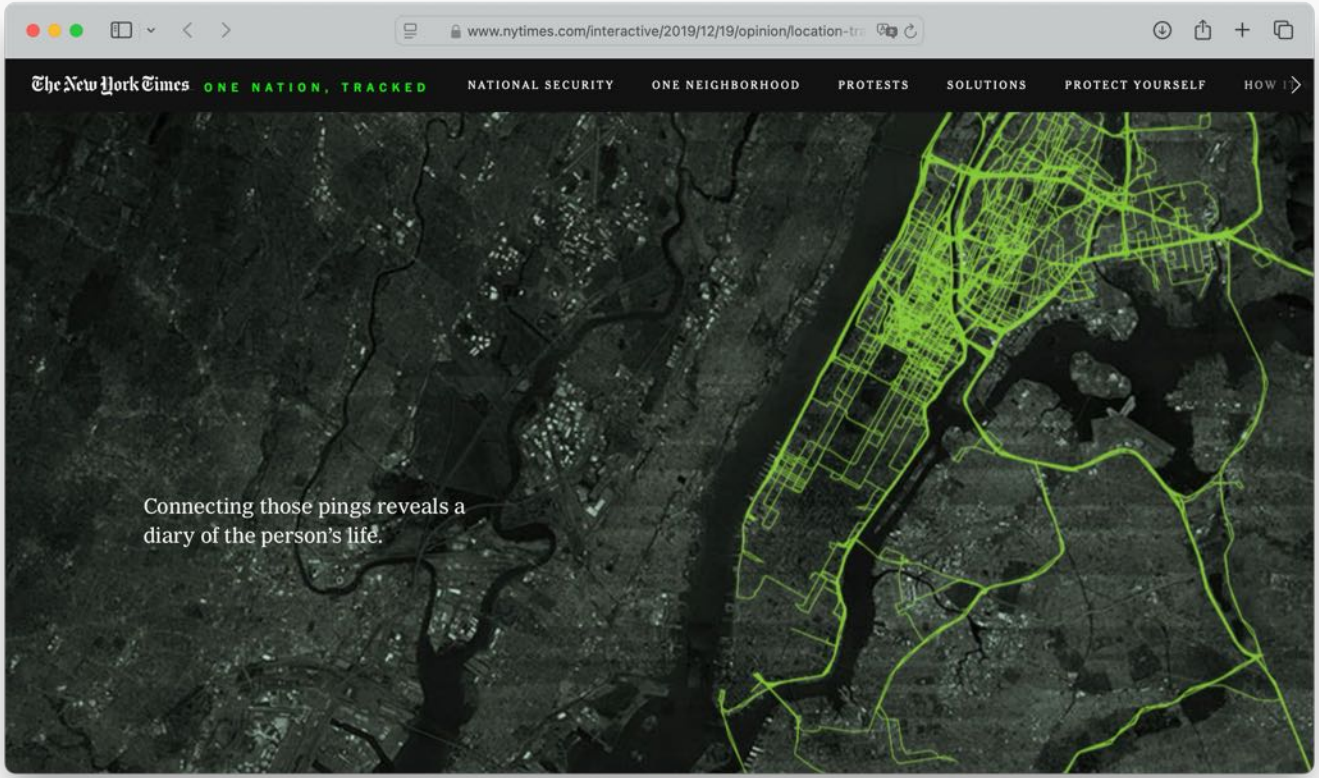


Drop-off locations for trips starting at Larry Flynt's Hustler Club between midnight and 6 am during 2013.

Source: http://content.research.neustar.biz/blog/differential-privacy/stripRaw.html (2014)

ONE NATION, TRACKED

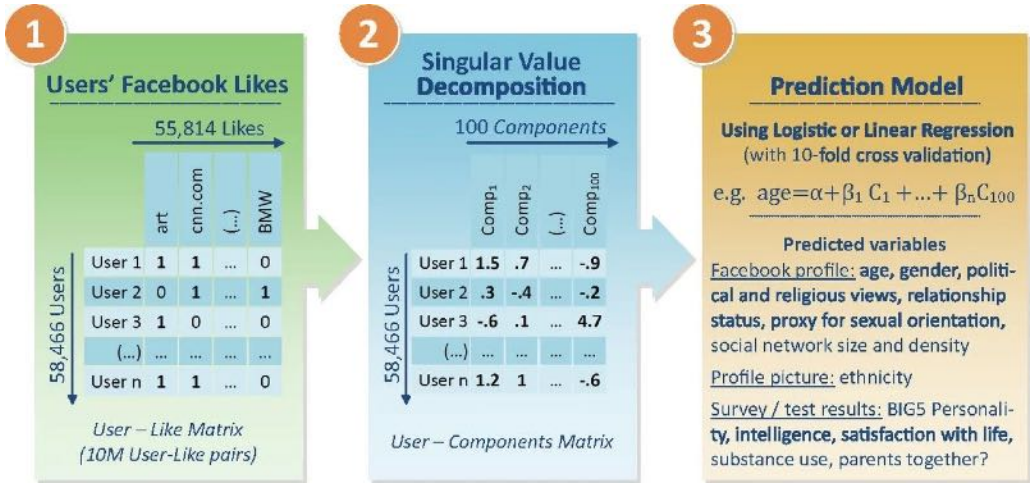
https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html



Predictive data analytics

»Facebook Likes can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.«

»The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases.«

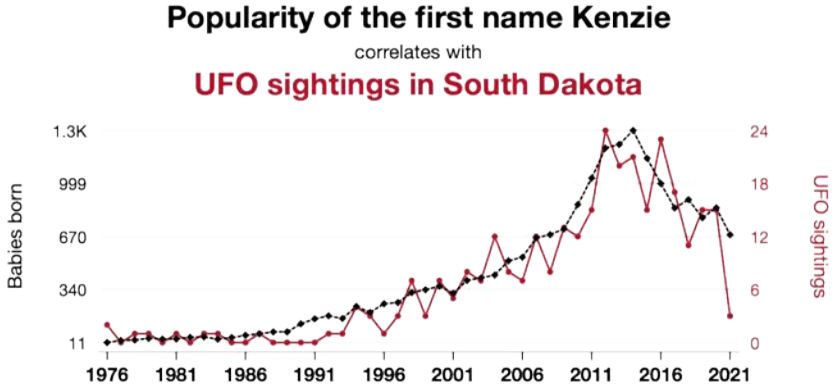


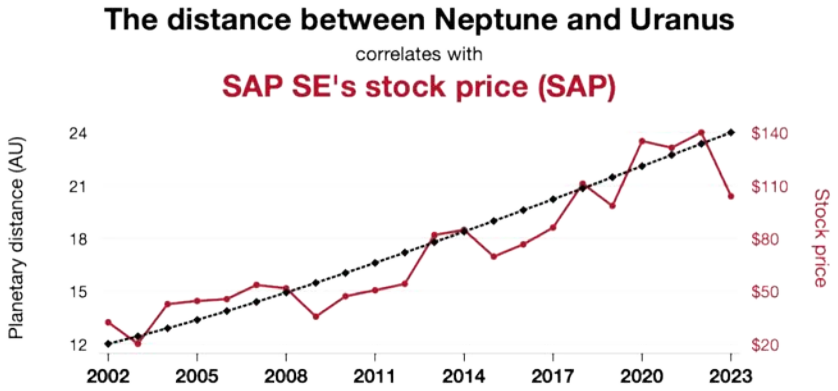
https://www.pnas.org/doi/full/10.1073/pnas.1218772110

Source and further reading: Michal Kosinski, David Stillwell, Thore Graepel: Private traits and attributes are predictable from digital records of human behavior. Proc. Nat. Academy of Sciences (PNAS) 110/15 (2013), https://doi.org/10.1073/pnas.1218772110

Correlation is not causation

https://tylervigen.com/spurious-scholar
https://www.tylervigen.com/spurious-correlations





spurious scholar

Because if $p < 0.05$, why not publish?

Step 1: Gather a bunch of data. Note

Step 2: Dredge that data to find random correlations between variables. Note

Step 3: Calculate the correlation coefficient, confidence interval, and *p*-value to see if the connection is statistically significant. Note

Step 4: If it is, have a large language model draft a research paper.

Step 5: Remind everyone that **these papers are AI-generated and are not real.** Seriously, just pick one and read the lit review section. Note

Step 6: ...publish:

Neubestimmung des Begriffs personenbezogene Daten

Nach: Hoffmann-Riem 2018, S.55

- Personenbezogene Daten sind auch Daten, die als Ergebnis einer Big-Data-Analyse entstehen.
  - allgemein und ohne Herleitung aus Daten speziell der konkret betroffenen Person
  - Beispiele: Person wohnt in einem bestimmten Stadtteil; daraus Ableitung von Finanzkraft, Herkunft, sexueller Orientierung, Gesundheit
- Personenbezogene Daten sind auch Daten, deren Personenbezug durch Anonymisierung entfällt.
  - Möglichkeiten der Deanonymisierung und Ableitung von Eigenschaften dürften nicht unterschätzt werden
  - Beispiele: New York Taxi Data Analytics, Strava Heatmap
- ebenso kritisch pseudonymisierte, aggregierte, perturbierte, verschlüsselte Daten betrachten

Verkehrszeichenerkennung mittels maschineller Lernverfahren

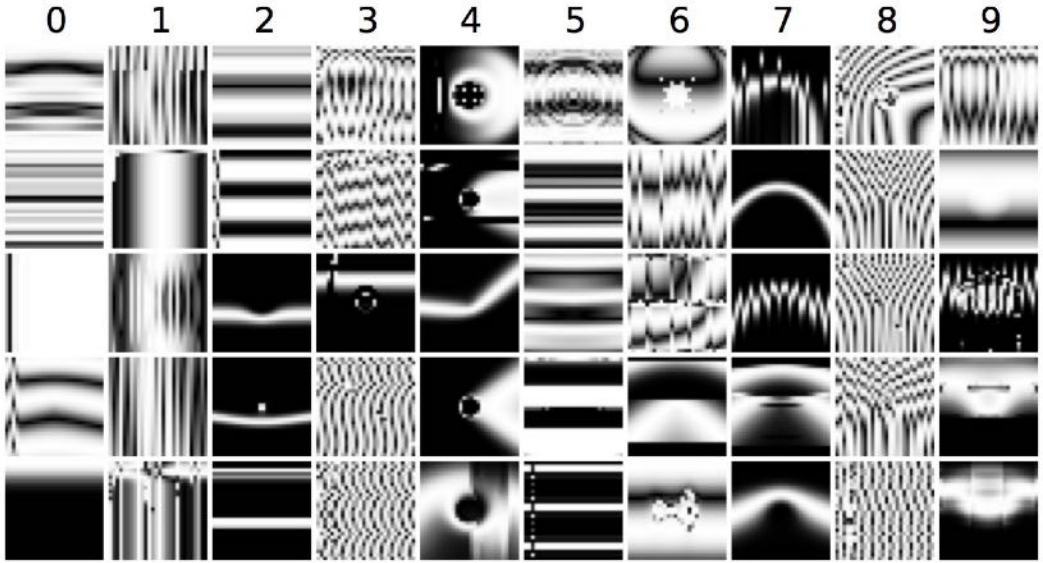
REALITÄT	ERKENNUNG: Gefahr	ERKENNUNG: »normal«
Gefahr	richtig positiv	falsch negativ
»normal«	falsch positiv	richtig negativ

J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural Networks, Volume 32, August 2012, Pages 323-332, ISSN 0893-6080

…klappt leider nicht immer

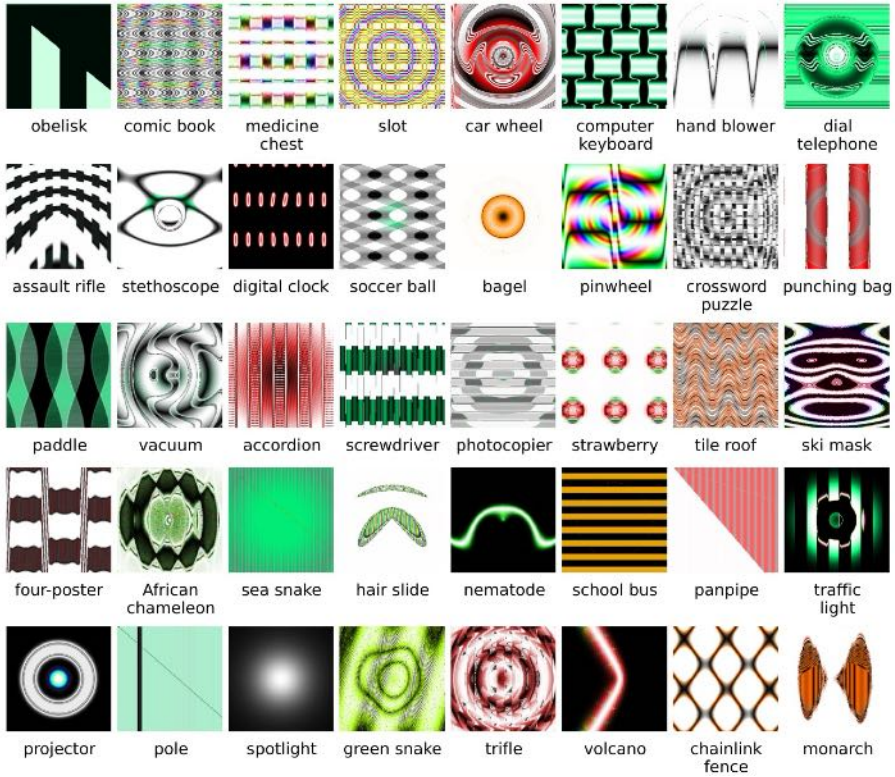
Erkannte Ziffer

Eingabedaten, die fälschlicherweise als Ziffer erkannt werden



Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In CVPR ’15, IEEE, 2015.

Gezielte Manipulationen eines autonomen Fahrzeugs möglich



Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In CVPR ’15, IEEE, 2015.

Verbergen von Eigenschaften beim Klassifizieren

- Adversarial learning

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

inf.uni-hamburg.de

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DEPARTMENT OF INFORMATICS
SECURITY AND PRIVACY

HOME COURSES THESES RESEARCH PEOPLE SERVICE

UHH → MIN-Fakultät → Fachbereich Informatik → Einrichtungen → Arbeitsbereiche → Security and Privacy → Home

WORKING GROUP ON «SECURITY AND PRIVACY»

Security and Privacy

Information systems become more and more important in critical infrastructures, while the Internet has evolved to a critical infrastructure itself. The secure operation of these infrastructures is vital and their failure can have severe impacts up to the loss of human lives.
Security refers to the fact that protection goals are achieved in the presence of malicious attacks and system failures. Typical security goals can be confidentiality, integrity, accountability, and availability. Security and privacy in information systems addresses both technical and organizational aspects, such as building and establishing security concepts and security infrastructures as well as risk analysis and risk management.
Privacy can be a conflicting goal to security, but they can also benefit from each other. Hence, it is necessary to balance both when developing secure information systems.

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

Telefon +49 40 42883 2358

hannes.federrath@uni-hamburg.de

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