



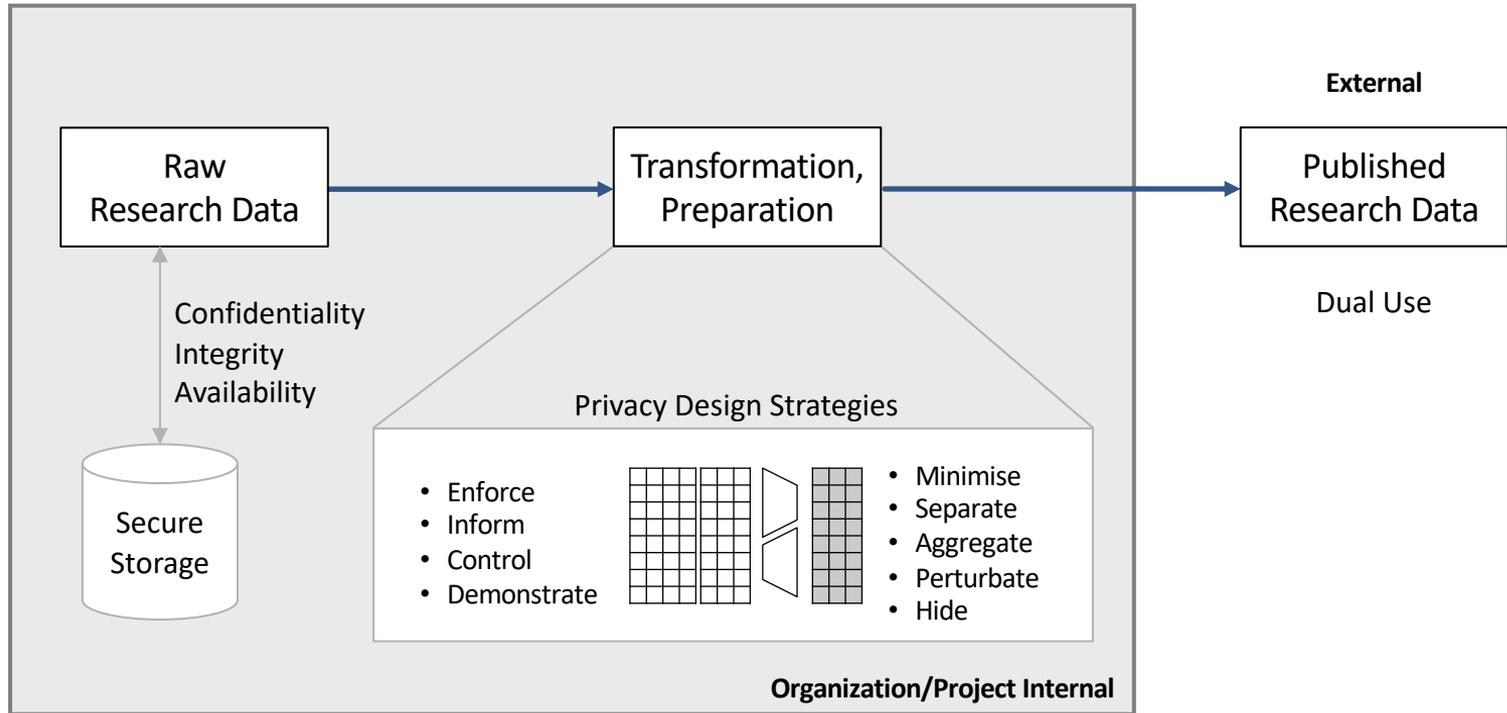
Datenschutzwahrende Methoden der Forschungsdatenverarbeitung

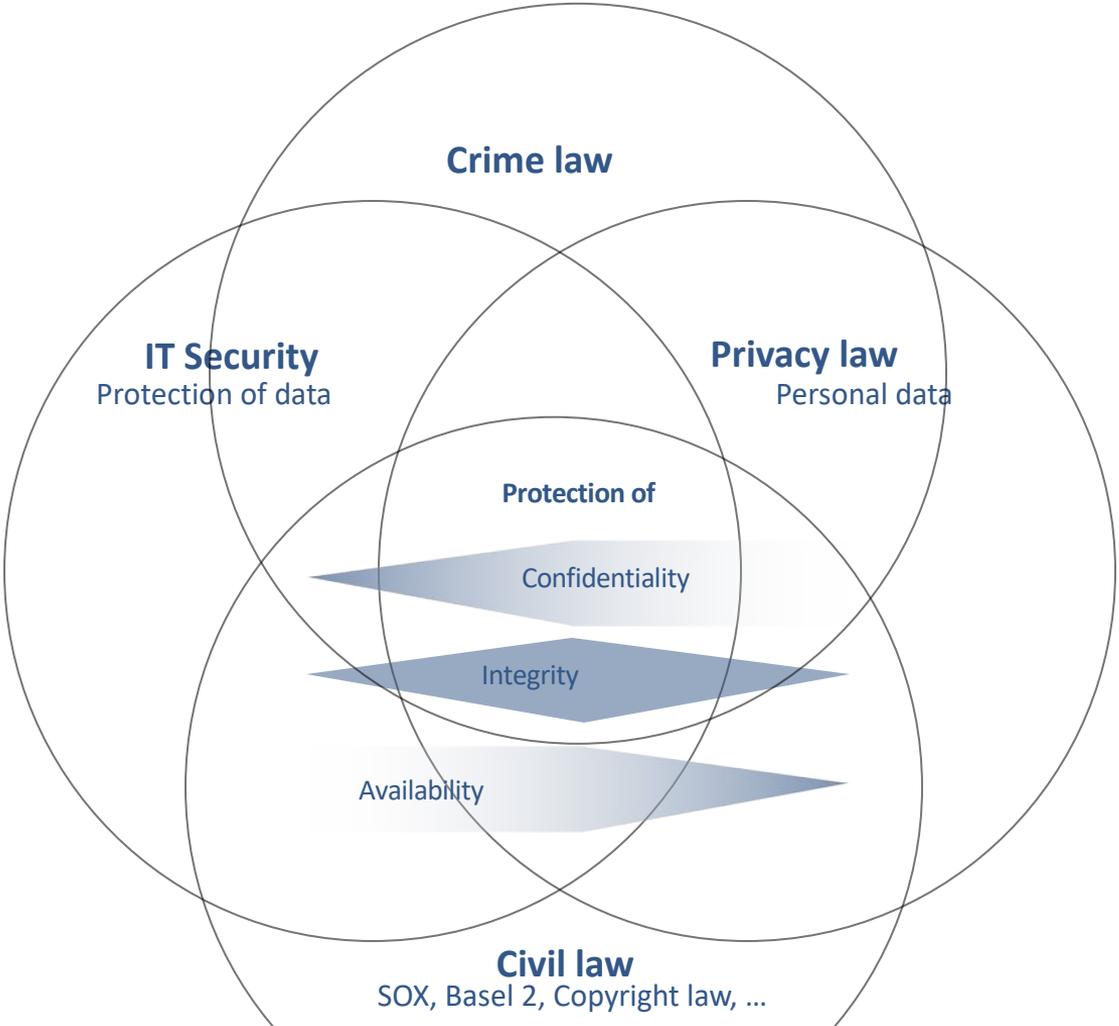
Prof. Dr. Hannes Federrath

Sicherheit in verteilten Systemen (SVS)

<http://svs.informatik.uni-hamburg.de>

Data Transformation process





Datenverarbeitung allgemein vs. Forschungsdatenverarbeitung

■ Forschungsdatenverarbeitung

– kurzfristig

- schnelle, konkrete Fragestellungen
- meist lokale Speicherung
- keine dauerhafte Speicherung

– langfristig

- Archivierung in Forschungsdatenbanken und Registern
- notwendige Daten oft zum Zeitpunkt des Aufbaus der Forschungsdatenbank noch nicht bekannt
- offene, bisher unbekannte Fragestellungen
- Verknüpfung von Daten aus verschiedenen Datenbeständen

Erkennbarer Zielkonflikt zwischen Nützlichkeit der Daten und Erforderlichkeit, Vereinbarkeit von Forschungsdatenmanagement mit den Datenschutzgrundsätzen (Art. 5 DSGVO) fraglich

■ Art. 5 DSGVO

(1) a) Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz

b) Zweckbindung

»Die Zwecke, zu denen personenbezogene Daten verarbeitet werden, sollten eindeutig und rechtmäßig sein und zum Zeitpunkt der Erhebung der personenbezogenen Daten feststehen.« (vgl. ErwG 39). Gesetzlich erlaubte Weiterverarbeitung u.a. für wissenschaftliche Forschungszwecke und statistische Zwecke wird »nicht als unvereinbar« angesehen

c) Datenminimierung

»auf das für die Zwecke notwendige Maß« beschränkt

d) Richtigkeit

unrichtige Daten unverzüglich löschen oder berichtigen

e) Speicherbegrenzung

»nur so lange [...], wie es für die Zwecke [...] erforderlich ist«; auf das erforderliche Mindestmaß begrenzen (ErwG 39)

f) Integrität und Vertraulichkeit

Datensicherheit (Art. 32)

(2) Rechenschaftspflicht

»Der Verantwortliche ist für die Einhaltung des Absatzes 1 verantwortlich und muss dessen Einhaltung nachweisen können«

Sicherheit der Verarbeitung

- Art. 32 (1) fordert in Präzisierung von Art. 25 (1)
 - geeignete technisch-organisatorische Maßnahmen
 - zur Pseudonymisierung und Verschlüsselung
 - zur Sicherstellung von Vertraulichkeit, Integrität, Verfügbarkeit, Belastbarkeit
 - zur Wiederherstellung der Verfügbarkeit nach Zwischenfällen
 - zur Überprüfung, Bewertung und Evaluierung der technisch-organisatorischen Maßnahmen



Typische Frage: Was ist die Rechtsgrundlage der Datenverarbeitung?

Sicht der Datenverwender

- **Datennutzung zur eigenen Forschung**
 - Was ist bei Veröffentlichung von Forschungsergebnissen?
- **Nutzung im Team**
- **Nutzung organisations- oder konzernweit**
- **Datenweitergabe an externe Forscher und Forscherinnen**

Privacy Impact

Blick auf Art. 6 DSGVO hilft...

- a) Einwilligung (Freiwilligkeit)
- ...
- e) öffentliches Interesse (gesetzliche Regelung)
- ...
- f) berechtigtes Interesse (Abwägung)

Blick der Betroffenen

- Datenspende, Selbstlosigkeit
- Kommerzielle Verwendung der Daten

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

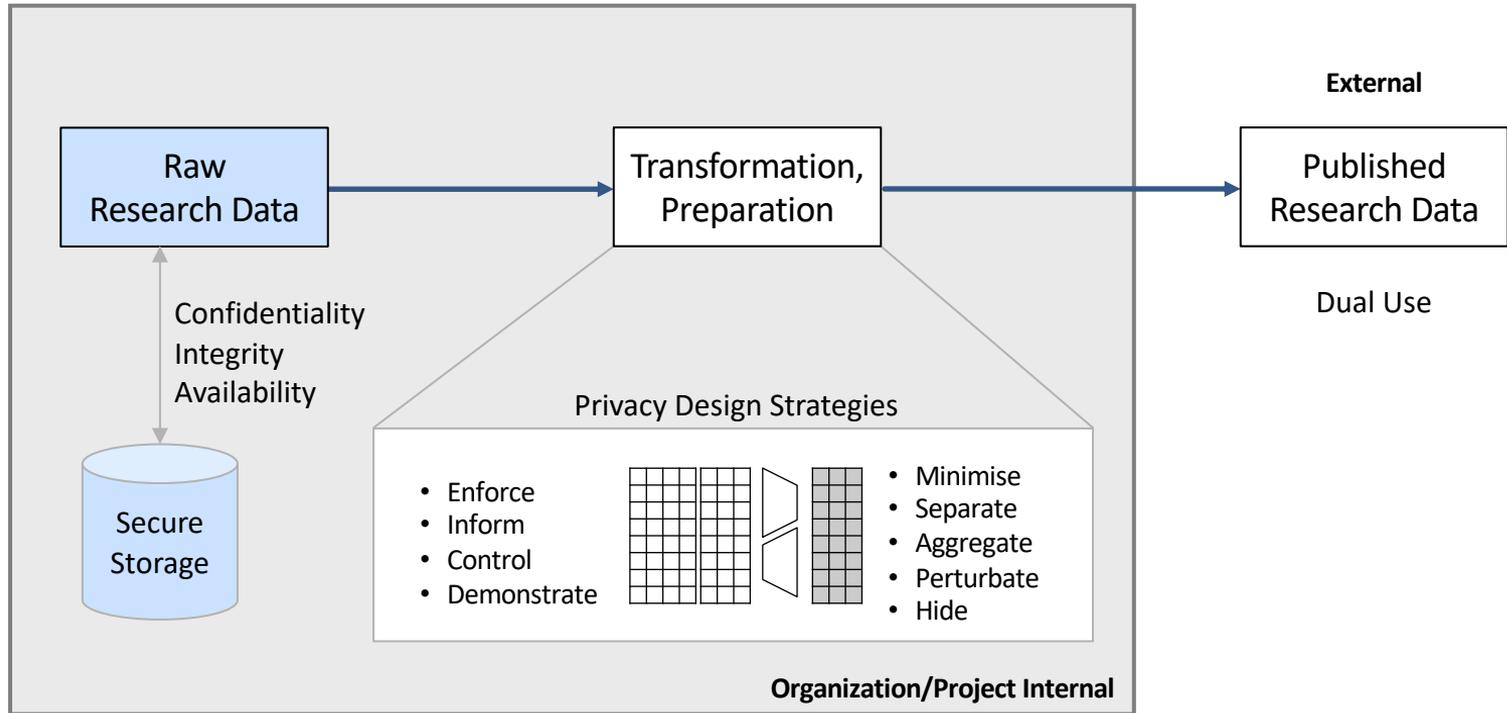
Es existiert kein milderes Mittel gleicher Eignung, den Zweck zu erreichen.



Angemessen

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

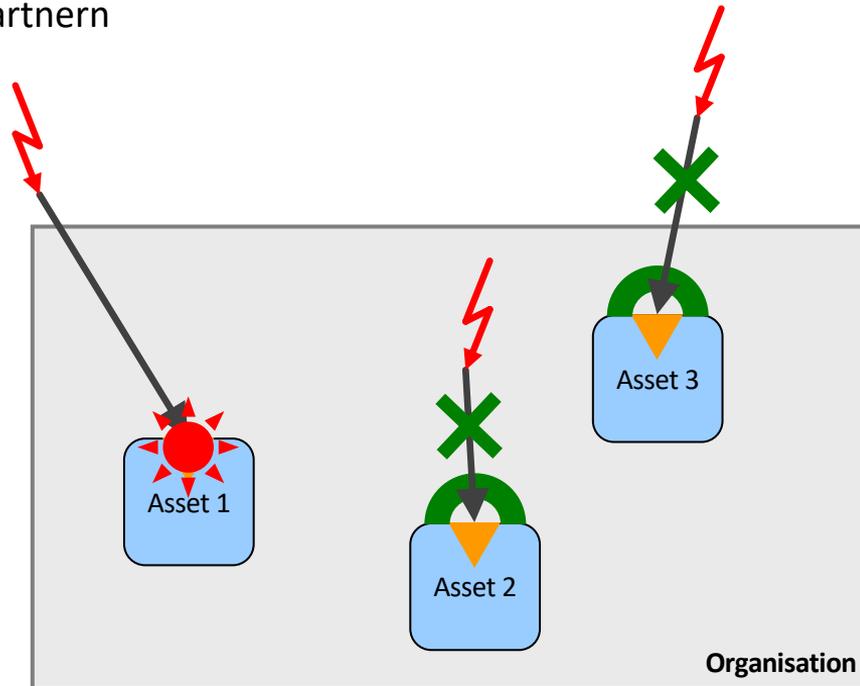
Data Transformation process



Von der Bedrohung zum Sicherheitsvorfall

■ Warum IT-Sicherheitsmanagement?

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



Bedrohungen, z.B.

- Viren, Würmer
- DoS
- Hacking
- Spionage
- Social Engineering

Verwundbarkeiten, z.B.

- Konfigurationsfehler
- Buffer Overflows

Schutzziele

- Vertraulichkeit
- Integrität
- Verfügbarkeit

Maßnahmen

- Präventiv
- Detektiv
- Reaktiv

vgl. Nowey, 2011

Vertrauen schaffende Maßnahmen als Schlüssel für Akzeptanz

- Hier zunächst kaum Unterschiede zur »normalen« Datenverarbeitung
 - strikte Transparenzgebote und wirksame Informationspflichten
 - gesetzliche Anforderungen und Mindeststandards zur Datensicherheit
- Aber: Drei Beispiele (Spezialfälle) aus Sicht der Betroffenenrechte, die teilweise leer laufen
 - Widerruf einer Einwilligung:
 - Rücknahme einer Datenweitergabe ist kaum möglich
 - Recht auf Korrektur:
 - Weiterverarbeitung (»Veredelung«) der Daten meist durch *Verknüpfung* mit anderen Daten
 - z.B. Anlernen von KI-Systemen
 - Rückführung und Veränderung nicht möglich
 - schlimmer noch: Durchsetzung des Rechts würde personenbeziehbare Daten erfordern
 - Informationspflichten: Wer wird die Daten erhalten?
 - nur unvollständig erfüllbar, da zum Zeitpunkt der Datenerhebung/Einwilligung unklar

Vertrauen schaffende Maßnahmen als Schlüssel für Akzeptanz

■ Lösungsansatz

1.

- Forschungsdatenverarbeitung sollte sich von Anfang an am Ziel ausrichten, so wenig wie möglich personenbezogene Daten zu verarbeiten, weil voller und reiner Personenbezug meist unnötig ist.

■ Forschung orientiert sich im Wesentlichen an dem Ziel, Strukturbildung zu betreiben und neue Phänomene zu verstehen.

- Notwendigkeit präziser Daten über Individuen, jedoch unmittelbarer Personenbezug meist unnötig

■ Folgerung

2.

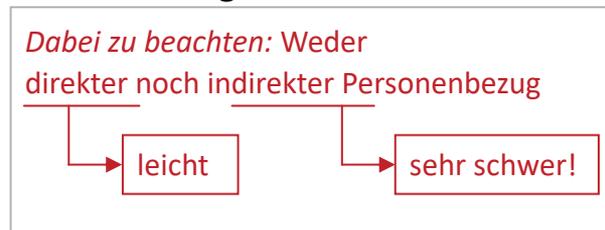
- Forschungsdatenverarbeitung sollte von den technischen Möglichkeiten der Anonymisierung und Pseudonymisierung reichlich und konsequent Gebrauch machen.

- Alle Daten sind, soweit dies technisch möglich und zumutbar ist, vor der Datenweitergabe zu Forschungszwecken bzw. der Speicherung in Forschungsdatenbanken so zu reduzieren und zu verändern, dass kein *Rückschluss auf die Identität einer natürlichen Person* möglich ist.

■ Grenzen des Schutzes

3.

- Forschungsdatenverarbeitung sollte mit dem klaren Bekenntnis betrieben werden, dass ein perfekter Schutz niemals möglich ist.



Indirekt identifizierbare natürliche Person

Daten sind auch dann personenbezogen, wenn die Zuordnung zu einer Person nur indirekt vorgenommen werden kann.

- entscheidend ist allein die (realistische) Möglichkeit einer Zuordnung
 - Zuordnung über Gruppenzugehörigkeit genügt bereits
 - Zuordnung auf Umwegen (Zusatzwissen, Big Data) ist ebenfalls eine realistische Möglichkeit
- Urteil des EuGH v. 19.10.2016, Az. C-582/14
- Es genügt auch, über rechtliche Mittel zu verfügen, die eine verantwortliche Stelle in die Lage versetzen, die Zuordnung zu einer Person vorzunehmen.
 - Hintergrund war: Personenbezug von dynamischen IP-Adressen für Webseitenbetreiber

Indirekt identifizierbare natürliche Person

■ Beispiel: Personenbezug von IP-Adressen

- entscheidend ist der Aufwand zur Herstellung des Personenbezugs
- wenn fehlender Personenbezug:
 - keine Anwendbarkeit des Datenschutzrechts
 - dürften unbeschränkt gespeichert und übermittelt werden
- Zulässigkeit der Speicherung: Zur Aufrechterhaltung des Dienstes
 - 14 Tage sind zu lang und damit nicht angemessen

■ Urteil des EuGH v. 19.10.2016, Az. C-582/14

- Es genügt auch, über rechtliche Mittel zu verfügen, die eine verantwortliche Stelle in die Lage versetzen, die Zuordnung zu einer Person vorzunehmen.
 - Hintergrund war: Personenbezug von dynamischen IP-Adressen für Webseitenbetreiber

- Auszug aus Erwägungsgrund 26 der DSGVO:

»Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern.

Bei der Feststellung, ob Mittel nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind. ...«

Anonyme Daten

- **Auszug aus Erwägungsgrund 26 der DSGVO:**

»Die Grundsätze des Datenschutzes sollten [daher] nicht für anonyme Informationen gelten, d. h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann. Diese Verordnung betrifft somit nicht die Verarbeitung solcher anonymer Daten, auch für statistische oder für Forschungszwecke.«

- **Absolute Anonymität**

- Daten können unter keinen Umständen mehr zugeordnet werden

- **Faktische Anonymität**

- Einzelangaben können nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden

- Datenschutzrecht grundsätzlich nicht anwendbar
- Durch Leistungssteigerung von IT-Verfahren können faktisch anonyme Daten wieder personenbezogen werden!

Pseudonymisierte Daten

- Art. 4 Nr. 5 DSGVO Pseudonymisierung

»Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden.«

Maßnahmen, die darauf abzielen, eine direkte Zuordnung (ohne Kenntnis einer Zuordnungsregel) zu einem Betroffenen zu unterbinden, ohne dabei in jedem Fall den Personenbezug dabei völlig aufzuheben

Pseudonymisierte Daten

- Art. 4 Nr. 5 DSGVO Pseudonymisierung

»Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden.«

- Auszug aus Erwägungsgrund 26 der DSGVO:

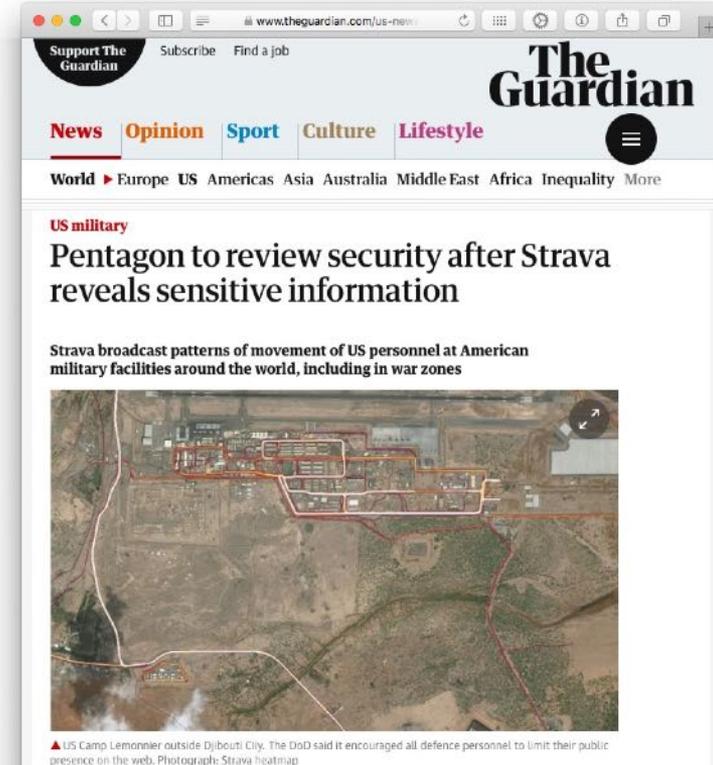
»Einer Pseudonymisierung unterzogene personenbezogene Daten, die durch Heranziehung zusätzlicher Informationen einer natürlichen Person zugeordnet werden könnten, sollten als Informationen über eine identifizierbare natürliche Person betrachtet werden.«

3. ■ Forschungsdatenverarbeitung sollte mit dem klaren Bekenntnis betrieben werden, dass ein *perfekter Schutz niemals möglich ist*.
- Beispiele:
- Der Fall »Strava Heatmap« (2018)
 - Auch Gruppeneigenschaften sind identifizierbare personenbezogene Daten
 - Analyse von 170 Mio. Open-Data-Datensätzen der Taxifahrten der NY Taxigesellschaft (2014)
 - Adressen der Interessenten von Nachtclubbesuchen wurden bekannt
 - SpiegelMining: Reverse Engineering von über 70.000 Spiegel-Online-Artikeln (2016)
 - Arbeits-, Urlaubs und Redaktionsgewohnheiten waren aus Daten analysierbar
 - eigene Untersuchungen zur Kontrolle und Überwachung der Arbeitsleistung von Softwareentwicklern (nächtliches Arbeiten, Produktivitätseinbrüche, Zusammenarbeitsstrukturen im Team) und von Daten aus privaten Haushalten legen Lebensgewohnheiten offen

Realität hat gezeigt, dass es immer wieder möglich war, Personenbezug herzustellen (zu re-identifizieren), weil die Möglichkeiten der Anonymisierung und Pseudonymisierung zu blauäugig angewendet wurden

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>

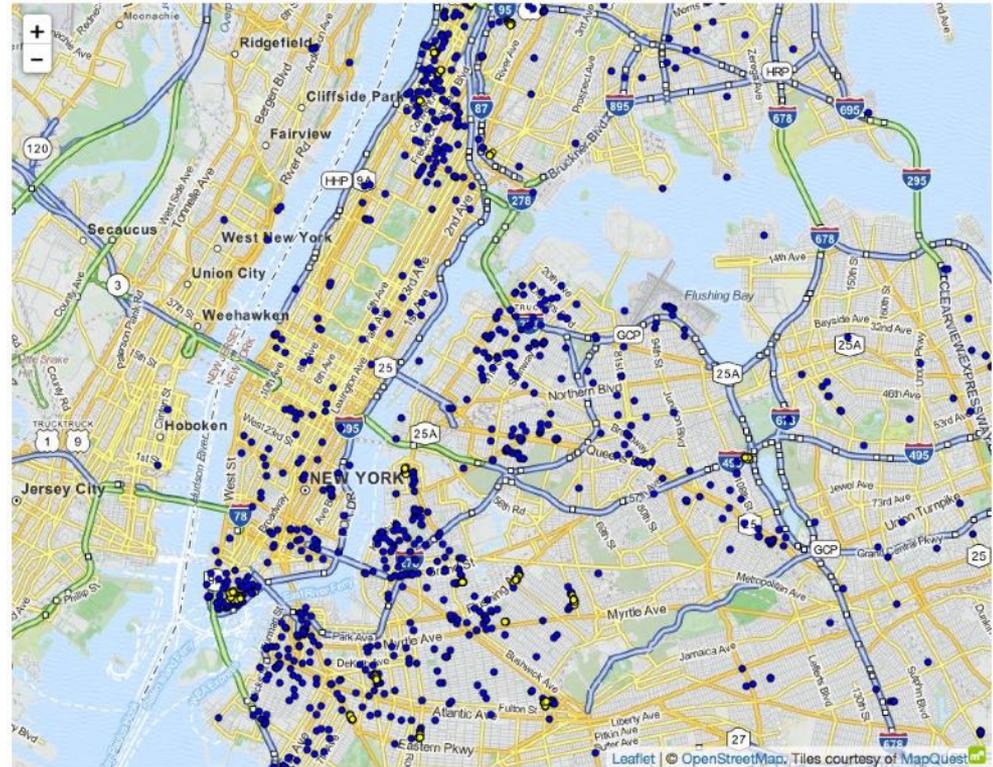
Pseudonymisierte Daten...

...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/stipRaw.html> (2014)

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.

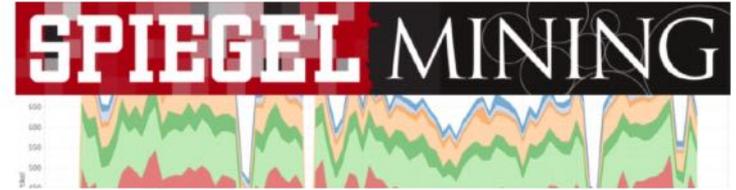
It also includes potential off-site meetups by factoring in "coincidental drop-offs," where someone from the New York Fed and someone from a bank each appeared to be dropped off at the same location around the same time.

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.

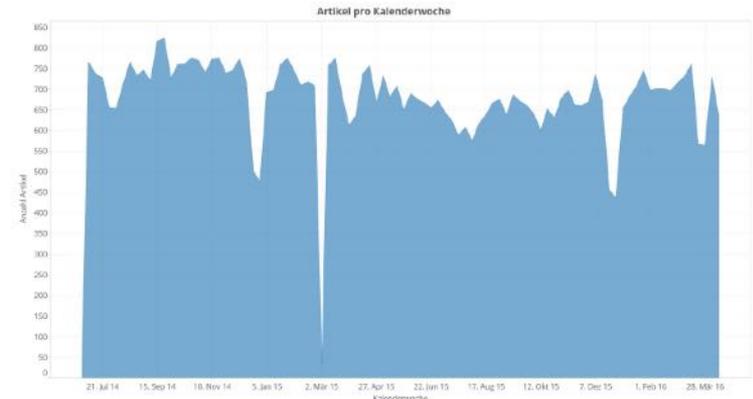
- Reverse Engineering von über 70.000 Spiegel-Online-Artikeln (2016)
- Verteilung von veröffentlichten Artikeln der Rubriken über das Jahr
- Kombination von Merkmalen: Textlänge und Zeit
 - längste Artikel wochentags zwischen 5 und 6 Uhr wochentags
 - längste Artikel am Wochenende zwischen 7 und 9 Uhr
- Auch Arbeits-, Urlaubs und Redaktionsgewohnheiten waren aus den Daten analysierbar.

https://www.dkriesel.com/blog/2016/0725_spiegelmining_analyse_70000_spiegelonline_artikel



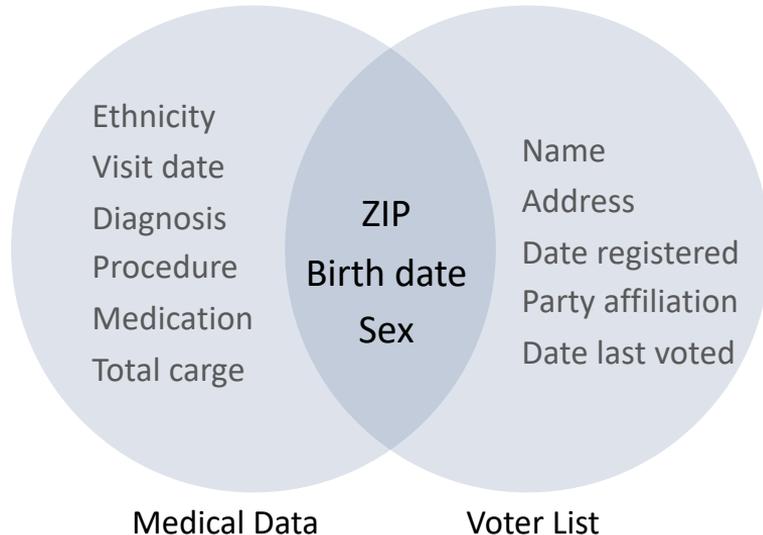
SpiegelMining. Auch Spiegelredakteure feiern Weihnachten. Eine Analyse von 70.000 SpiegelOnline-Artikeln

Seit Mitte 2014 habe ich mehr als 70.000 Artikel von SpiegelOnline systematisch gespeichert. Jeden Tag kommen im Schnitt 100 dazu. Diese Artikelmasse werden wir in der nächsten Zeit auswerten und erforschen. Was herauskommt, ist eine tiefgreifende Analyse des Publikationsverhaltens des vielleicht größten Meinungsmachers Deutschlands.



Vermeintlich anonymes
medizinisches Register mit
Daten von US-Bürgern...

...wurde verknüpft mit
öffentlich zugänglichen US-
Wählerverzeichnissen



Beide Datensätze enthalten Geschlecht, Geburtsdatum, Postleitzahl.

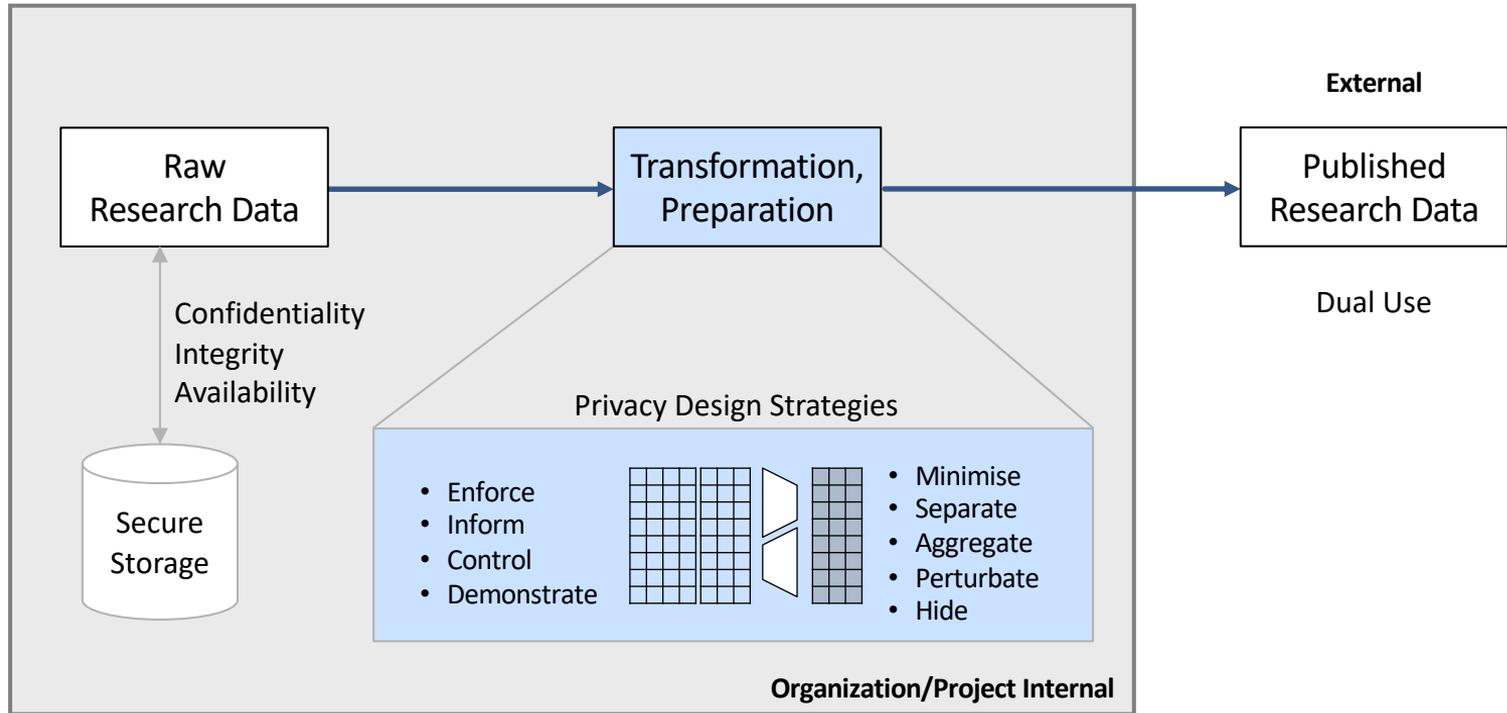
Ergebnisse:

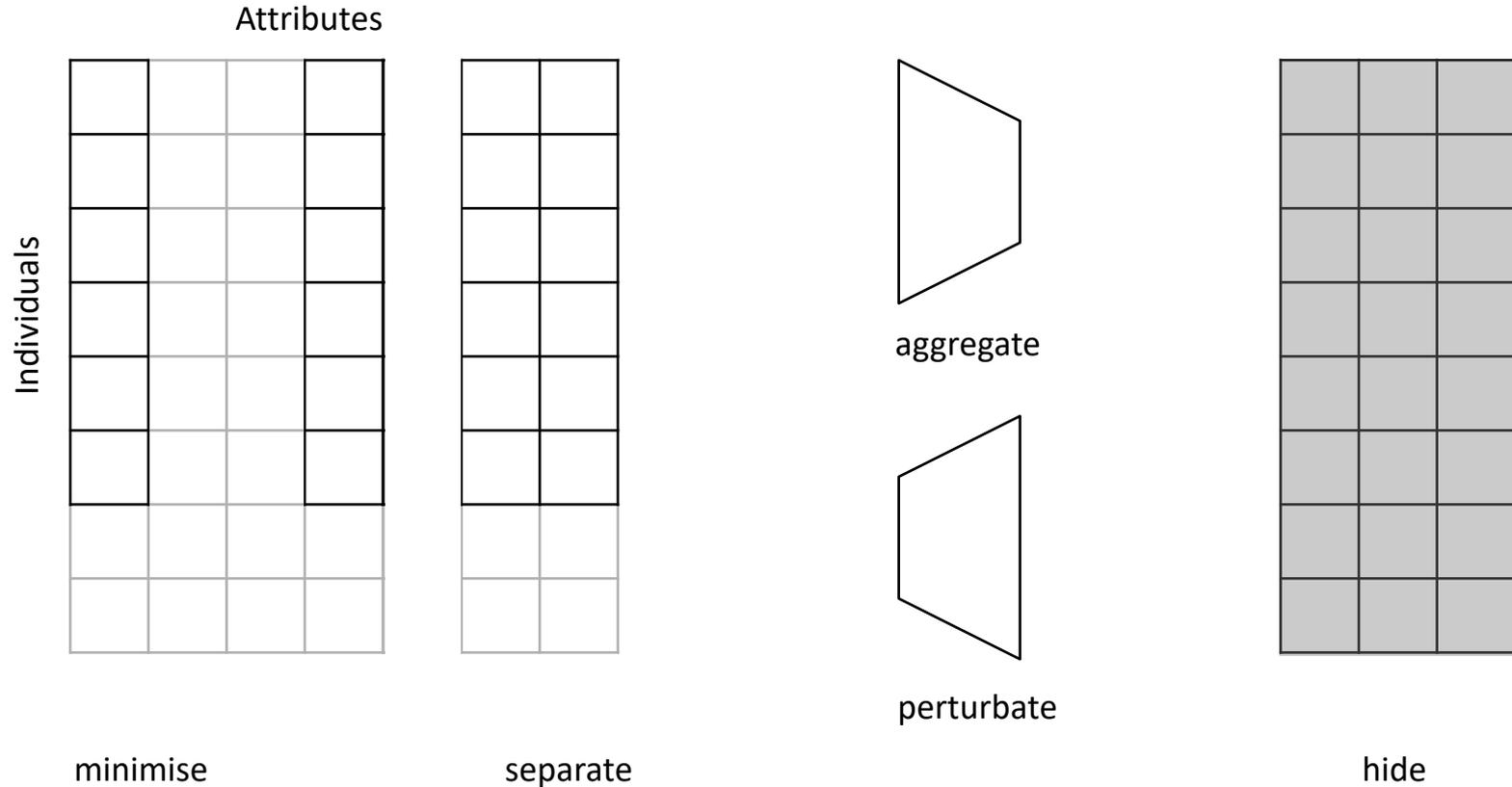
- Identifizierung der Krankenakte des ehem. Gouverneurs von Massachusetts, William Weld, war möglich
- Insgesamt 87 Prozent der US-Bevölkerung kann re-identifiziert werden



Latanya Sweeney entwickelte das Konzept der k-Anonymität.

Data Transformation process





■ 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

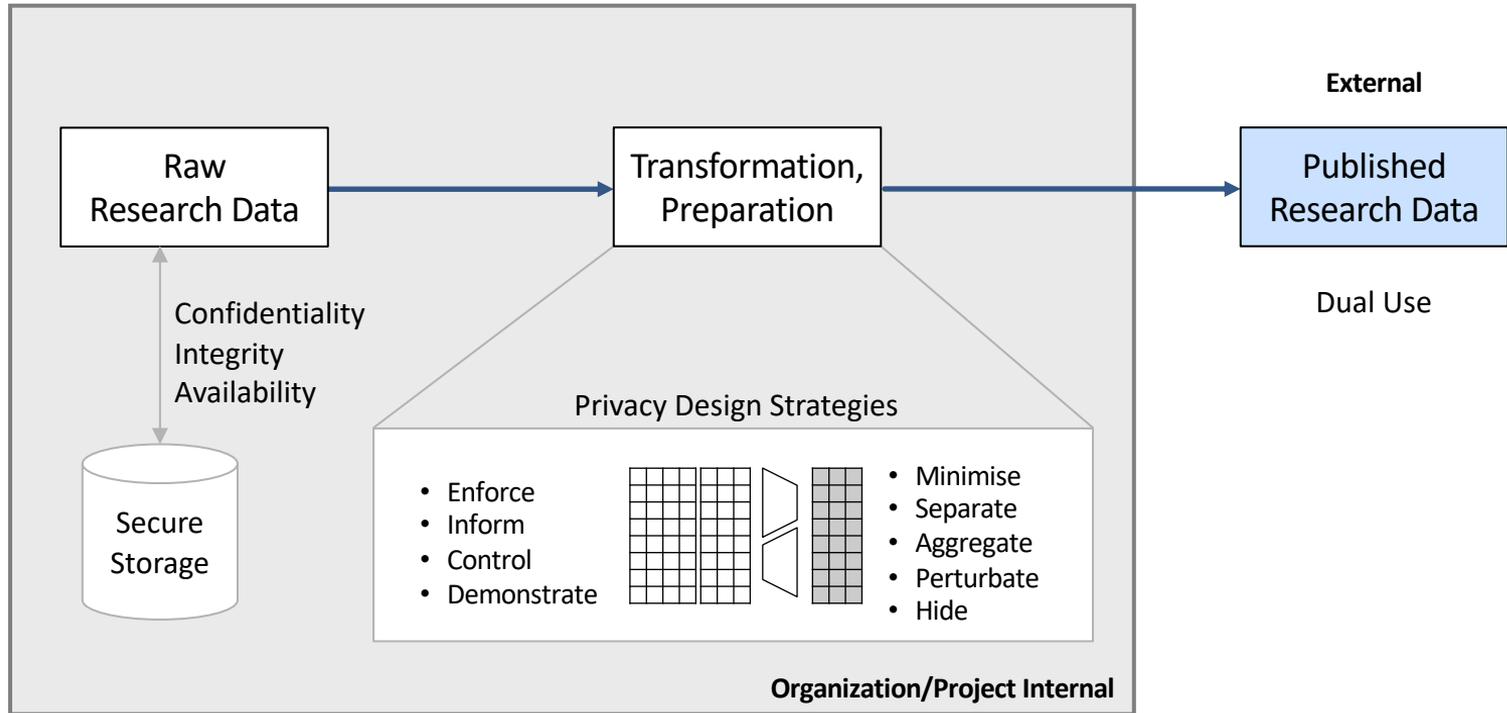
- **Enforce**: Durchsetzung einer Datenschutz-Policy (access control)
- **Inform**: Betroffene über Datenverwendung informieren (P3P)
- **Control**: Eingriffsmöglichkeit der Betroffenen (informed consent)
- **Demonstrate**: Überprüfbarkeit (privacy management, logging)

Offenheit ist aus technischer Sicht der Weg zum Ziel einer Transparenz im Sinne der Teilnehmerüberprüfbarkeit.

- **Schlüsseigenschaften offener Systeme**
 1. klar definierte Schnittstellen zum Zugriff
 2. quelloffene Softwaresysteme
 3. Offenlegung und damit gute Dokumentation des Systemdesigns

- **Wege zu einer Vertrauensbasis**
 - Forderung einer Datenschutzfolgenabschätzung (DSFA)
 - mindestens bei gesetzlicher Befugnis zur Forschungsdatenverarbeitung
 - bei der Verwendung der Daten in KI-Systemen, etwa zum Anlernen von solchen Systemen

Data Transformation process



- **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
 - Trends und 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

Prädiktive Big-Data-Analytik: Stecknadeln im Heuhaufen



- 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 Deanononymisierung und Ableitung von Eigenschaften dürften nicht unterschätzt werden
 - Beispiele: New York Taxi Data Analytics, Strava Heatmap
- ebenso kritisch pseudonymisierte, aggregierte, perturbierte, verschlüsselte Daten betrachten



Notwendigkeit forschungsethischer Grundsätze

■ Governance

- Kaskade von berufsspezifischen Regelungen
- bisher kaum existent für Forschungsdatenverarbeitung

Berufsgeheimnis

Berufsethische Grundsätze

Forschungsethische Grundsätze

■ Beispiel Deutsche Forschungsgemeinschaft (DFG):

- Leitlinien zum Umgang mit Forschungsdaten
 - Projektplanung und Antragstellung
 - Bereitstellung der Daten in sinnvoller Weise (Rohdaten nicht notwendigerweise)
 - langfristige Sicherung (mind. 10 Jahre)
- Wissenschaftliche Fachgemeinschaften sollen an Regelungen arbeiten
 - Disziplinspezifische Regelungen
 - Anerkennung der Leistung bei der Verfügbarmachung von Forschungsdaten

Gesetzliche Regelungen

Open Data ist die verschärfte Form von Forschungsdatenmanagement
Risiken sind unüberschaubar

- Daher
 - Nutzen von synthetisch generierten Daten anstelle von personenbeziehbaren Daten
 - Nutzung von Verfahren zum verteilten, förderierten Analysieren von Daten
- Grundidee
 - Forscherinnen und Forscher sollen an der Datenquelle analysieren, Strukturen bilden

inf.uni-hamburg.de

 **Universität Hamburg**
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DEPARTMENT OF INFORMATICS
SECURITY AND PRIVACY

[HOME](#) [COURSES](#) [THESES](#) [RESEARCH](#) [PEOPLE](#) [SERVICE](#) 



SECURITY AND PRIVACY

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>