Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Vorlesung

Datenschutz und Privatheit in vernetzten Informationssystemen

Kapitel 2: Digitale Identitäten

Erik Buchmann
buchmann@ipd.uka.de

Agenda für heute

- *(Wiederholung: Prinzipien des Datenschutzes)*
- Digitale Identitäten
- Bedrohungen durch die digitale Identität
- Größte Gefahr: Identitätsdiebstahl
- Verkettung digitaler Identitäten
- Fallbeispiel: Verkettung in Suchmaschinen

→ *Nächste Vorlesung:*
Entkettung mittels Anonymisierung

Wiederholung: Prinzipien des Datenschutzes

- Jeder Bürger soll selbst bestimmen können, und
- Jeder Bürger soll wissen,
  - wer was wann und unter welchen Bedingungen
    - über ihn weiß.
    - über ihn in Erfahrung bringen darf.

- Ausnahmen nur auf gesetzlicher Basis
  - wenn das Interesse Dritter bzw. der Allgemeinheit schwerer wiegt als die Schutzinteressen des Betroffenen

Klassische Datenschutz-Schwerpunkte

- Schutz des Bürgers vor dem Staat
  - Recht auf freie Meinungsäußerung wird bedeutungslos, wenn Regierung den Sprecher im nachhinein identifizieren (und abstrafen) kann
    → **Datenschutz wichtig für Demokratie**

- Schutz des Bürgers vor privaten Unternehmen
  - Sind die individuellen Vorlieben und Absichten des Käufers bekannt, wird perfekte Preisdifferenzierung und Manipulation des Käufers möglich
    → **Datenschutz ist Kundenschutz**

Neue Schwerpunkte der letzten Jahre

- Schutz des Bürgers vor dem Dienstanbieter
  - Verknüpfung und Mining von personenbezogenen Daten, die als Nebenwirkung moderner Dienste anfallen, erlaubt unerwartete Rückschlüsse *(Beispiel: Google erhebt keine Daten über Nutzer, sondern Nutzer geben diese über Suchbegriffe preis.)* → **Verbleib persönlicher Daten nachvollziehen**

- Schutz des Bürgers vor dem Bürger
  - Werden in Online-Communities private Details und Beziehungen öffentlich preisgegeben, lassen sich Persönlichkeitsprofile von Unbeteiligten erstellen. → **Kontrolle über persönliche Daten behalten**

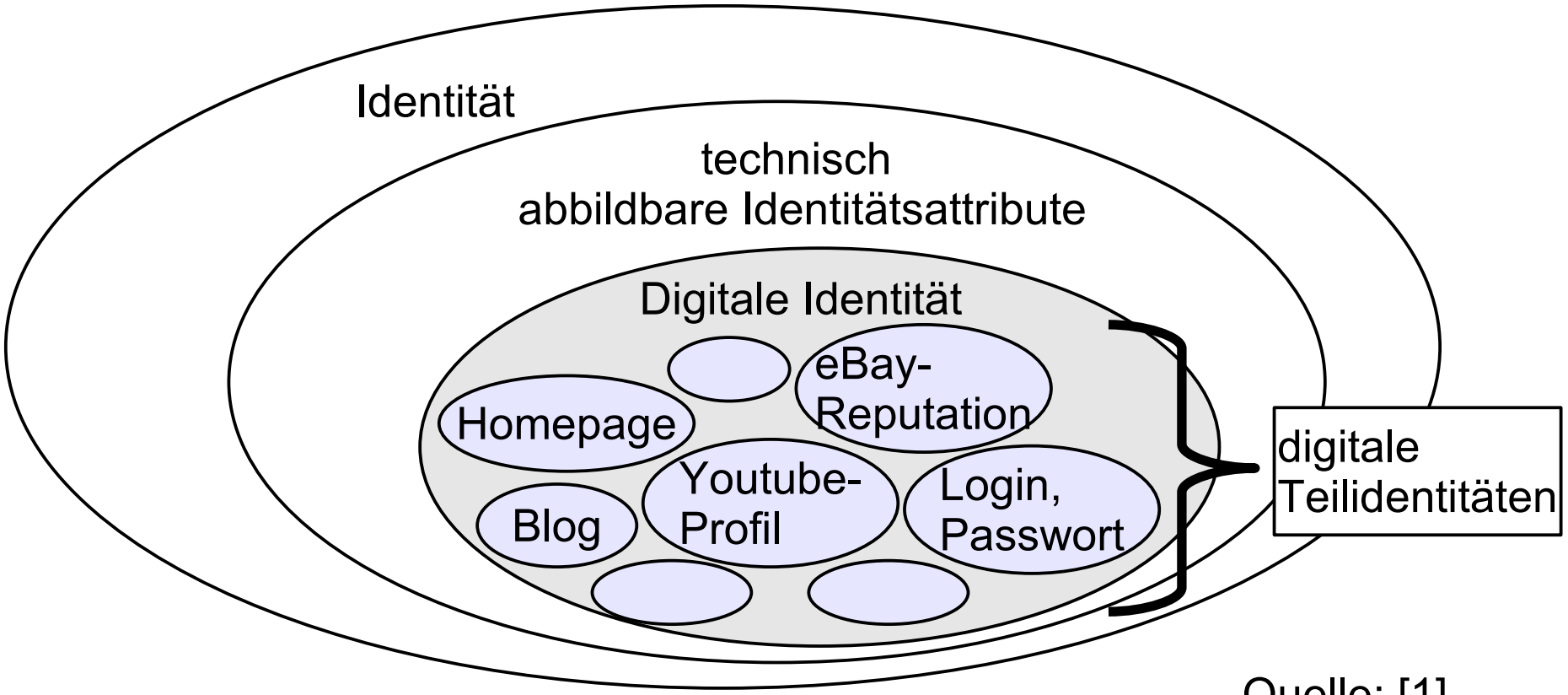
Die Digitale Identität

Die Digitale Identität

- Definition: *"Jede mögliche Form von technisch abgebildeten Daten, die zu einer Person gehören" [1]*
  - Daten zur eindeutigen Authentifizierung, z.B. Adresse, Name, biometrische Daten
  - Daten zur pseudonymen Identifizierung, z.B. Login, Passwörter, Nicknames, Foren-Namen
  - Persönliche Merkmale, z.B. Vorlieben, Hobbies, Religion, Lebensumfeld
  - nicht unbedingt *von jedem* einer Person zuordnbar
    - Beispiel: IP-Adresse ist Teil der digitalen Identität, aber nur vom Internet-Provider zuordnbar

Übersicht: Identität im Netz

- Digitale Teilidentität: Untermenge der digitalen Identität, die eine Untermenge der abbildbaren Attribute sind
  - viele separate **digitale Teilidentitäten** möglich



Quelle: [1]

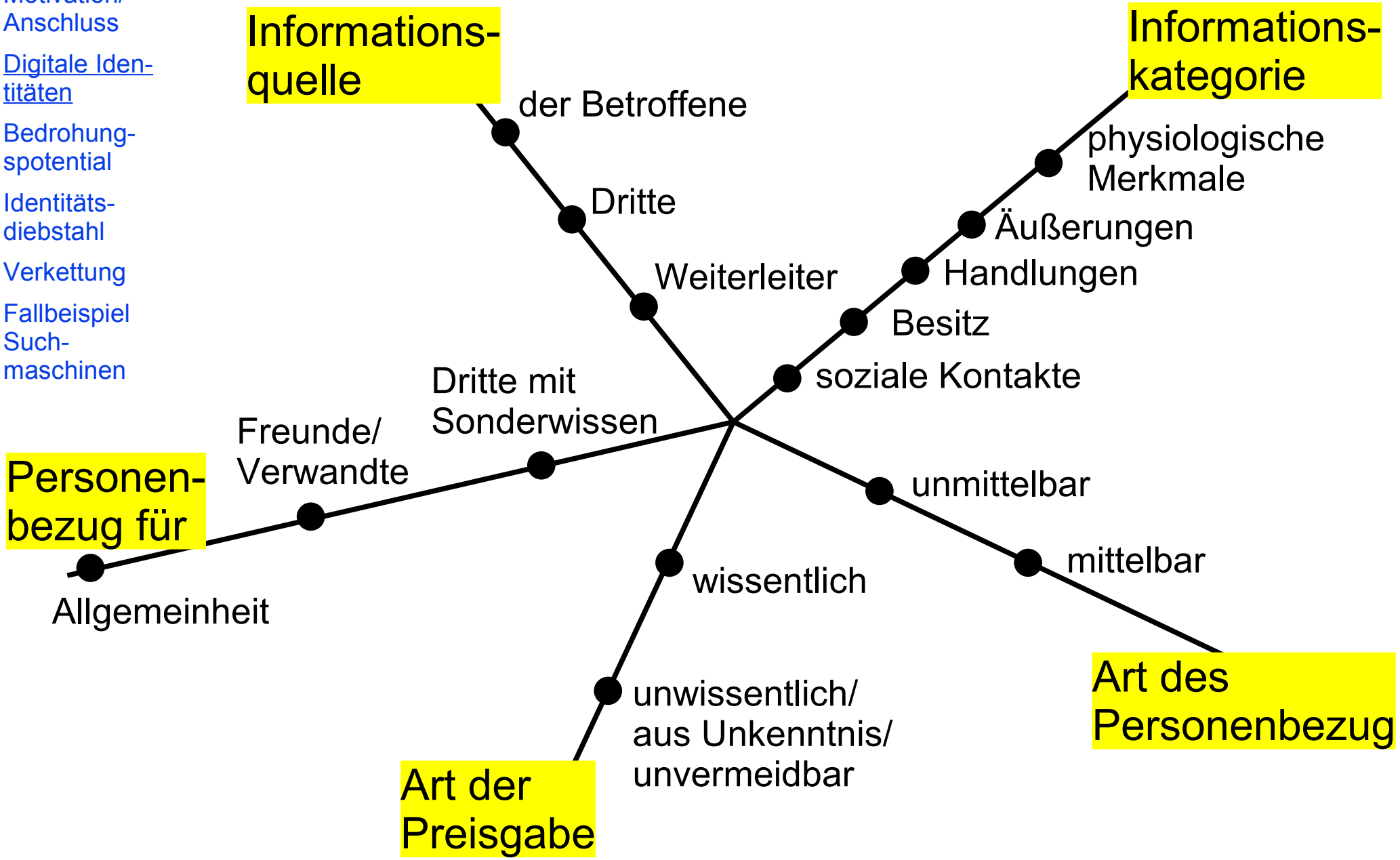
Unterschiedliche digitale Teilidentitäten

- Datenspuren im täglichen Leben
  - Blogs, Soziale-Netzwerk-Portale, Internet-Communities
  - Finanzamt, Wählerlisten, Anträge bei Behörden
  - Verkehr mit Wirtschaftsunternehmen, Einkäufe im Supermarkt oder Web-Shops

- Nicht alle Datenspuren werden wissentlich hinterlassen
  - Was weiß der Supermarkt (Rabatt-System) oder der Mobilfunkprovider (Positionsdaten vom Handy)?

→ durch **Verkettung von Teilidentitäten** Aufbau umfassender Persönlichkeitsprofile möglich!

Persönliche Daten: 5 relevante Dimensionen



Informationsquellen, unmittelbarer Bezug zum Betroffenen

	Informationsquelle: der Betroffene selbst	Informationsquelle: Dritte
Physiologische oder genetische Merkmale	z.B. gemessen von Sensoren: Größe, Gewicht, Augenfarbe, DNA, Fingerabdruck	z.B. Untersuchungen von Verwandten mit ähnlichen biometrischen Merkmalen
Äußerungen	z.B. Einträge in Web-Formulare, eigene Homepage, Profile in Online-Communities	z.B. Äußerungen über den Betroffenen, Bewertung in einem Reputationssystem
Handlungen	z.B. etwas kaufen, an einer Veranstaltung teilnehmen	z.B. im Namen anderer Personen tätig sein
Soziale Kontakte	z.B. Adressbuch-Einträge im Mailclient, Skype-Kontaktliste	z.B. veröffentlichte Kontakte in Online-Communities
Besitz	z.B. Kleinanzeigen, eBay-Angebote	z.B. im Auftrag des Betroffenen einkaufen

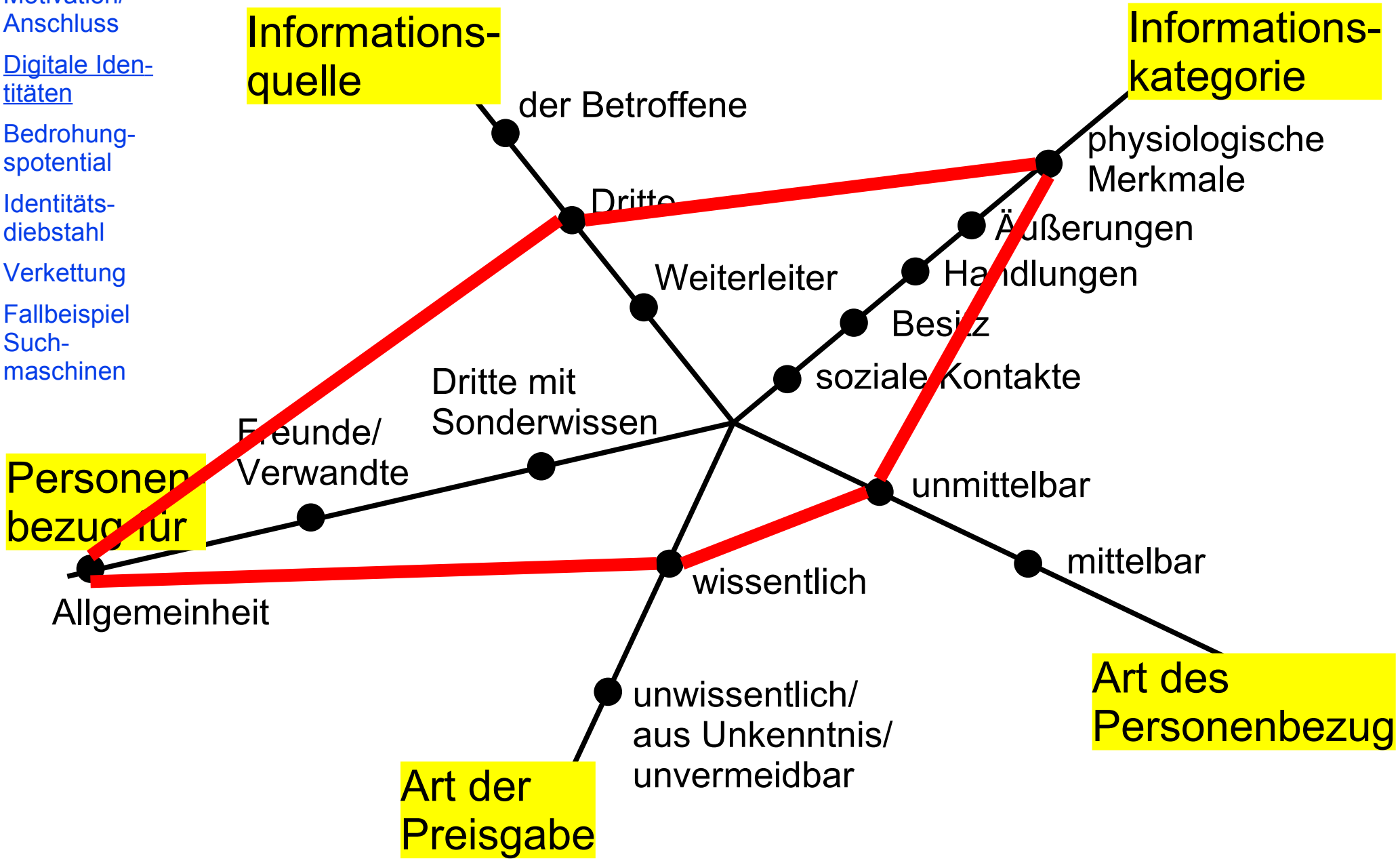
Quelle: [1], erweitert

Informationsquellen, mittelbarer Bezug zum Betroffenen

	Informationsquelle: Dritte
Äußerungen	z.B. Äußerungen über eine Personengruppe, der der Betroffene angehört
Handlungen	z.B. über Data Mining, Collaborative Filtering auf die Absichten des Betroffenen schließen
Indirekte Kontakte	z.B. Freund-meines-Freundes-Kontakte in Online-Communities
Besitz	z.B. Collaborative Filtering: Kundengruppe X kauft häufig Gegenstand Y

Quelle: [1], erweitert

Beispiel: Namentlich annotiertes Bild auf Flickr



Bedrohungen durch die digitale Identität

Probleme mit digitalen (Teil-)Identitäten

- Falschinformationen
- Datenmißbrauch
- Zweitverwertung
- Langfristige Aufbewahrung
- Verknüpfbarkeit
- Öffentliche Zugänglichmachung

(Anm.: Liste ohne Anspruch auf Vollständigkeit)

Falschinformationen

- Offensichtliches Problem: falsche Daten können
  - zu einem falschen Bild in der Öffentlichkeit führen
  - wirtschaftlich schädlich sein, z.B. falsche Schufa-Einträge, falsche Steuer-Berechnung
- Problem
  - Korrektur schwierig umzusetzen, wenn Daten vielfach redundant vorhanden sind
    - Daten in Backups
    - Usenet-Nachrichten, P2P-Dateien vielfach gespiegelt
  - Korrektur schwierig einzufordern, wenn Daten nicht gewerbsmäßig verarbeitet werden, sondern von privaten Anwendern (Wikipedia, Facebook)

Datenmißbrauch

- Personenbezogene Daten werden ohne Wissen oder Zustimmung des Betroffenen an Dritte übermittelt
  - einfachste Form: Handel mit Adresslisten
- Ausnahme: einige Arten der Datenübermittlung ohne Wissen und Zustimmung sind gesetzlich legitimiert, z.B.
  - Steuerfahndung
  - polizeiliche Ermittlungen
- Problem
  - unklar, wer Kenntnis von persönlichen Daten hat
  - Verbleib persönlicher Informationen für den Betroffenen nicht nachvollziehbar

Zweitverwertung

- Daten vom Geschäftsbetrieb für andere Zwecke nutzen
  - Data Mining auf Kundendaten
  - Autobahn-Maut (Toll Collect) → Strafverfolgung
  - Handy-Verbindungsdaten → Positionsbestimmung
  - Werbung
- Problem
  - Daten fallen oft unvermeidlich an, z.B. muss Lieferant die Lieferadresse kennen
  - Grenze zwischen legitimer Nutzung für den Geschäftszweck und illegitimer Nutzung oft unklar
  - für den Betroffenen aufgrund fehlender Transparenz oft nicht nachvollziehbar [2]

Verknüpfbarkeit

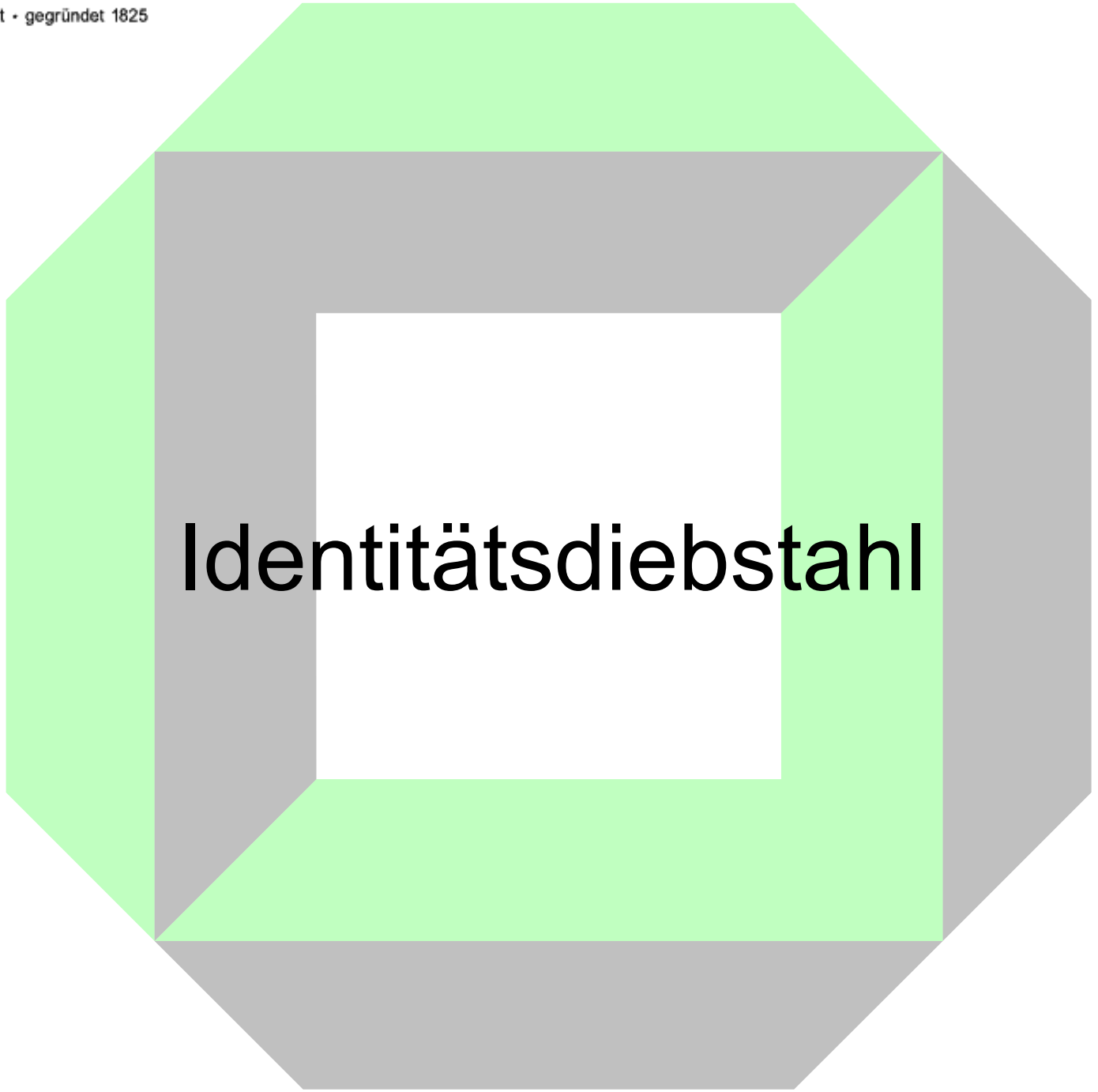
- Zusammenführen von “harmlosen” digitalen Teilidentitäten aus unterschiedlichen Quellen
  - öffentliche Quellen: Telefonbuch, Schufa-Daten, Handelsregister, Liegenschaftsbuch, etc.
  - nichtöffentlich: Daten aus dem Geschäftsbetrieb, Informationen von Behörden, Banken
- Anreichern von eigenen Daten mit Zusatzinformationen
- Suche nach diskriminierenden Merkmalskombinationen in verlinkten Daten → Rasterfahndung
- Problem
  - Aufbau von komplexen Persönlichkeitsprofilen
  - Fehler in den Daten können zu falschen Schlüssen führen

Langfristige Aufbewahrung

- Daten können beliebig lange gespeichert werden
- Rekonstruierbarkeit und Nachvollziehbarkeit von Aussagen oder Handlungen wird möglich
- Anmerkung: schon aus Praktikabilitätsgründen umfasst die Löschpflicht des BDSG keine Backups
- Problem
  - Selbstdarstellung in der Vergangenheit kann drastisch von aktuell gewünschter Selbstdarstellung abweichen
    - während der Schulzeit Partylöwe, jetzt Lehre zum Bankkaufmann
  - Kontrollverlust über einmal preisgegebene Daten; was einmal im Internet steht ist 'verbrannt'

Öffentliche Zugänglichmachung

- Auch objektive Fakten können zu einer falschen Selbstdarstellung führen
  - Auswahl an Informationen entscheidend, z.B. nur wenig schmeichelhafte oder veraltete Daten
- Selbst-Inszenierung des persönlichen Auftritts und Selbst-Bild sind wichtig, z.B.
  - soziale Kontakte, Selbstwertgefühl
  - Vorstellungsgespräche, Verkaufsgespräche
- Problem
  - auch persönliche Daten, die der Betroffene selbst veröffentlicht hat, sind schützenswert
  - Preisgabe von persönlichen Daten durch Dritte, z.B. Freundeslisten in Web-Communities


Identitätsdiebstahl

Mißbrauch personenbezogener Daten

- Personenbezogene Daten werden ohne Wissen oder Zustimmung des Betroffenen zu Zwecken gesammelt, gespeichert, verarbeitet oder übermittelt, die
  - den Interessen des Betroffenen zuwiderlaufen, z.B.
    - **Identitätsdiebstahl**
    - Stalking, Mobbing
    - Kreditbetrug
  - aber nicht gesetzlich legitimiert sind, z.B.
    - Steuerfahndung
    - polizeiliche Ermittlungen
- Für den Betroffenen nicht zu verhindern, da ohne Wissen und Zustimmung erfolgt

Identitätsdiebstahl

- *"A fraud committed or attempted using the identifying information of another person without authority."* (Definition der Federal Trade Commission, USA)
  - unlegitimierte Nutzung einer fremden Identität
  - betrügerischer Vermögensvorteil unter Inkaufnahme von Nachteilen für den "Inhaber" der Identität, Kreditgebern, Händlern etc.

- Zwei Spielarten:
  - **New Account Fraud**
  - **Account Takeover**

Hoofnagle, Chris Jay, *Identity Theft: Making the Known Unknowns Known.* Harvard Journal of Law and Technology, Vol. 21, 2007.

Identitätsdiebstahl USA

- Zahlen 2008: Data Breach Stats (*"Einbruchstatistik"*)

	# of Breaches	*# of Consumer Records*
Banking/Credit/Financial	78	18,731,947
Business	240	5,886,960
Educational	131	806,142
Government/Military	110	2,954,373
Medical/Healthcare	97	7,311,833
total:	**656**	**35,691,255**

- Quelle: http://www.idtheftcenter.org, Breach Database *(Anm.: Zahlen beschreiben nur Verlust von pers. Daten, nicht die Mißbrauchsfälle)*

Identitätsdiebstahl Deutschland

- Zahlen 2008
  - 17 Mio. Datensätze von T-Online-Kunden im Umlauf
    - Handynummern, Adressen, Geburtsdaten, E-Mail
  - Call-Center-Angestellte kopieren Daten von 30 Mio. Telekom-Kunden
  - Verbraucherschützer kauft 6 Mio. Datensätze von Bundesbürgern (incl. Kontoverbindung) für 850 EUR
  - Whistleblower findet 1.5 Mio. Datensätze der Süddeutschen Klassenlotterie in einem Call-Center
    - Name, Telefonnummern, Bankverbindung

New Account Fraud

- Generieren einer künstlichen Identität
  - echte Daten, um Validierungsverfahren zu täuschen
    - Kontoeröffnung USA: Social Security Number
    - Kontoeröffnung Deutschland: Name, Postanschrift für Schufa-Auskunft
  - künstliche Daten zum Vervollständigen
    - Geschlecht, Alter, Beruf, Einkommen, Familienstand
  - künstliche Daten, um Plausibilität zu erhöhen
    - Vermögenslage (Rechnungen, Hypotheken, Kreditkartenkonten etc.)
    - Lebenslauf

Erkennung ist schwierig

- Künstliche Identitäten bestehen aus Mischung von echten und falschen Daten
- Inhaber der echten Daten erfahren oftmals nur indirekt vom Mißbrauch
  - Korrespondenz, Unterlagen, Mahnungen gehen an den Betrüger
  - Schaden entsteht meist indirekt, z.B. wenn Schufa-Auskunft belastet
- Aus Händler- oder Bankensicht kein Unterschied zwischen flüchtigem Schuldner und gefälschter Identität
  - unklar, ob Betrugsfall oder gewöhnlicher Kreditausfall

Account Takeover

- Übernahme einer bestehenden digitalen Identität
  - *Pishing*, gefälschte EMails erfragen Kontodaten, eBay-Konten, Kreditkartennummern, *www.meinebank.de.pisher.org*
  - *Pharming*, Webbrowser wird durch DNS-Spoofing o.ä. auf manipulierte Webseiten umgeleitet, die eBay- oder Banken-Webseiten gleichen
  - *Malware* auf dem Rechner protokolliert Anmeldeinformationen
  - *Social Engineering*,
    beliebtes Passwort: Name der Freundin,
    beliebte PIN: Geburtsdatum des Kindes

Erkennbarkeit

- Account Takeover oft leichter zu erkennen als New Account Fraud
  - die Betroffenen erhalten zumeist Mahnungen, Rechnungen etc.
- gesetzlicher Schutz des Betroffenen
  - Rückbuchung von per Lastschrift eingezogenen Beträgen
  - Stornierung von Kreditkartenrechnungen
  - Strafanzeige gegen Unbekannt

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

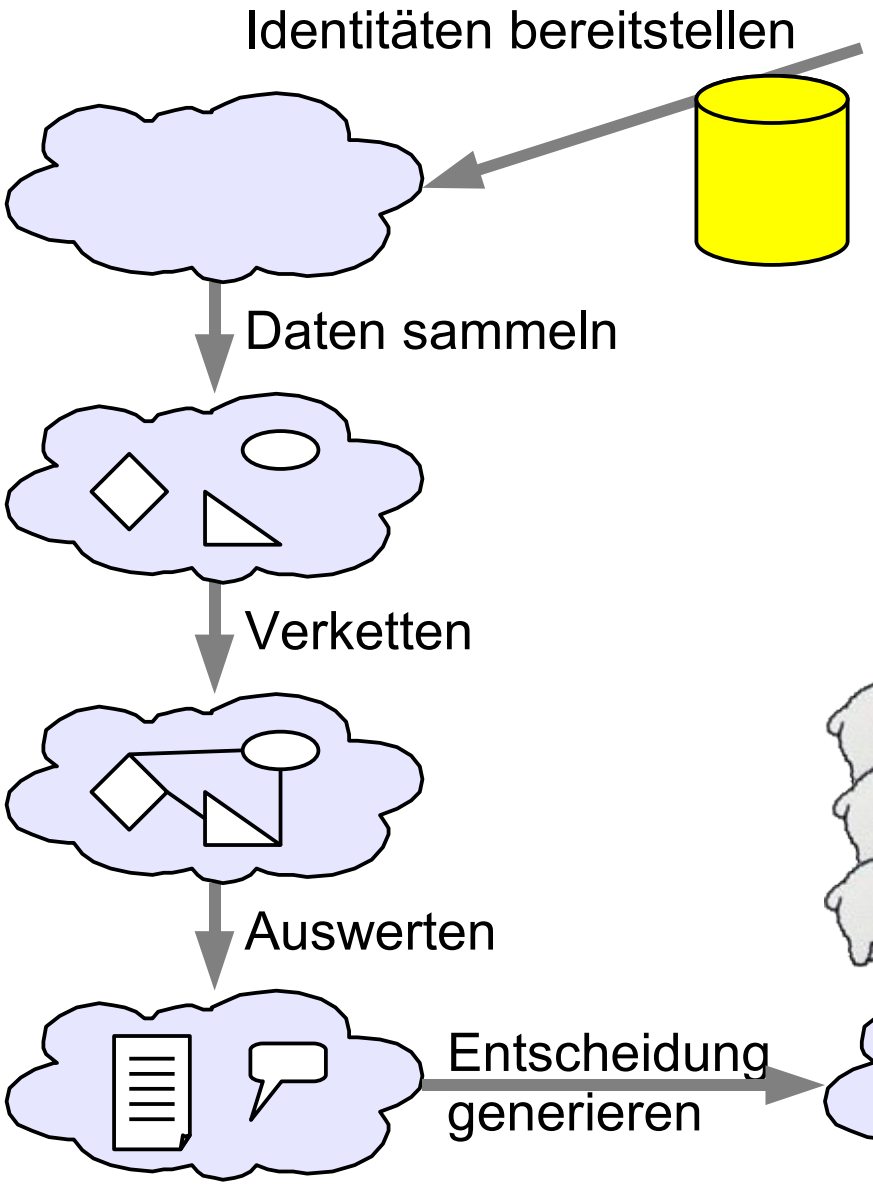
Verkettung Digitaler Identitäten

IPD, Forschungsbereich *Systeme der Informationsverwaltung*

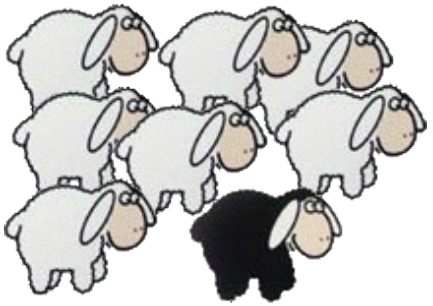
Anmerkung zum Begriff “Verkettung”

- Im Verlauf der Vorlesung Unterscheidung zwischen
  - **Verkettung** als Oberbegriff für das In-Beziehung-Setzen von digitalen (Teil-)Identitäten
    - Mengen von technisch abbildbaren Attributen mit Bezug zu einer Identität
  - **Verknüpfung** als Begriff für das technische Verbinden (Join) von Datenbeständen
    - Verbund über Primär-/Fremdschlüsselbeziehungen, Quasi-Identifier
    - bezieht sich auf das technische Verfahren, unabhängig vom Personenbezug
      → *siehe nächste Vorlesung*

Modell der Informationsanreicherung



- 5 aufeinander aufbauende Phasen



korrekt Betroffene

false positive,
false negative

Quelle: [1]

Phase 1: Vorbereitung

- Identitäten bereitstellen
  - Ausgangsmaterial für Folgephasen; bestimmt die Zielgruppe der Datenerhebung
  - insbesondere wichtig: Adress-Attribute, mit deren Hilfe Personen eindeutig ansprechbar sind
    - Name, Anschift, Telefonnummer
    - IP-Adresse, EMail-Adresse,
    - Login-Namen, Foren-Pseudonyme, ICQ-UIN, Skype-ID
- Adress-Attribute
  - temporär einer Person zuzuordnen
    - dynamische IP-Adresse, www.sofortmail.de
  - langfristig einer Person zuzuordnen
    - EMail-Adresse

Phase 2: Datensammlung

- Erheben und Speichern von Daten
  - durch den Betroffenen selber oder seine Kommunikationspartner
  - durch Ansammeln von “Datenspuren”, d.h. Logfiles, Sensordaten, etc.
- Typische Datenspuren heute:
  - *Standort* (IP-Adresse, Mobil- und Festnetztelefon, Lokationsbasierte Dienste, EC-Karte, Schlüsselkarte)
  - *Interessen* (besuchte WWW-Seiten, Suchmaschinen)
  - *Kommunikationspartner* (Telefon, ICQ, EMail, Skype)
  - *Bewegungsprofil* (DB-Fahrkartenautomat, Maut-Brücken, Überwachungskamera-Netzwerke)
  - *Anschaffungen* (Einkauf mit Rabattkarten, Stöbern auf Amazon.de, Anklicken von Werbebannern)

Phase 3: Verkettung

- Verkettungsalgorithmus kombiniert gesammelte Daten, z.B. mittels
  - Data Mining, Clustering, Association Rule Mining
  - Collaborative Filtering, Recommender-Algorithmen
  - Logfile-Analyse
  - Social-Network Analysis
- Verkettung üblicherweise zu einem vordefinierten Zweck
- Verkettung häufig anhand von Quasi-Identifiern
  - quasi-eindeutige Kombinationen von Identitätsmerkmalen als Schlüsselattribut
    → *nächste Vorlesung*

Phase 4: Auswertung

- Interpretation der verketteten Daten
  - Auswertealgorithmen, z.B.
    - unter Verwendung statistischer Modelle, um für den gegebenen Zweck relevante Profile mit einer gegebenen *Wahrscheinlichkeit* herauszufiltern
    - als Einzelfalluntersuchung, um relevante Provile mit hoher *Sicherheit* zu identifizieren
  - je intimer die gesammelten Daten, desto höher der mögliche Einfluss auf die Privatheit der Betroffenen

Phase 5: Generieren von Entscheidungen

- Entscheidungen mit Hilfe der Auswertungsergebnisse
  - Folgen für Individuen oder Gruppen von Betroffenen
    - Marketing; Verkaufsangebote und Reklame
    - Scoring; Kredite oder Hypothekenzinsen
    - Ranking; Job-Angebote
- Problem: Entscheidungen auf Basis ungenauer Daten
  - false positives: Betroffener wird fälschlich als Mitglied einer Gruppe identifiziert (z.B. Rasterfahndung)
  - false negatives: Betroffener wird fälschlich *nicht* als Gruppenmitglied identifiziert (z.B. kein Kredit)
  - Ranking, Scoring

Abhängigkeiten zwischen Phasen

- Fehler in den Daten pflanzen sich durch die Phasen fort
  - Falsche Datengrundlage
  - Fehler beim Sammeln der Daten; Zuordnung zur falschen Identität
  - Fehler in der Implementierung vom Verkettungsalgorithmus
  - Fehler im Auswertealgorithmus, Schwächen im Auswertemodell
  - Fehler bei der Entscheidung
  - Fehler bei der Datenübermittlung zwischen Phasen

Fehlentscheidungen

- Problem: Ausgestaltung der Phasen ist häufig Geschäftsgeheimnis; für Betroffenen intransparent
  - Schwer, sich gegen Fehlentscheidungen zu wehren *BDSG: Recht auf Auskunft, Sperrung, Löschung, Korrektur → eine der nächsten Vorlesungen*
- Aus Datenschutz-Sicht
  - falls möglich Daten gar nicht erst erheben
  - wenn erforderlich, dann Identitätsattribute wählen, die nur eingeschränkt verkettbar sind, z.B.:
    - Name, Adresse sind allgemeingültige Attribute, lassen sich beliebig verketten
    - Foren-Pseudonym nur in einem Web-Forum gültig, Verkettbarkeit stark eingeschränkt

Eigenschaften von Identitätsattributen

Weniger relevant für Privatsphäre	Potentiell gefährlich für die Privatheit
anonym	eindeutig identifizierend
nicht wiedererkennbar	wiedererkennbar
ändert sich im Verborgenen über die Zeit	unveränderlich
leicht änderbar	nicht änderbar
weitergebbar / übertragbar	nicht weitergebbar/übertragbar
flüchtig	langfristig gespeichert
nur einmal verwendet	häufig wiederverwendet
Authentizität unklar	authentisch / bestätigt durch Dritte
Zugriff anderer nicht möglich bzw. kontrollierbar	Zugriff anderer möglich oder intransparent
ermöglicht keinen direkten Kontakt	ermöglicht unmittelbaren Kontakt
unauffällig/geht in der Masse unter	unnormal oder herausragend
für wenige Teile des eigenen Lebens relevant / als trivial empfunden	betrifft zentrale Bereiche des täglichen Lebens / besonders sensibel
keine zusätzlichen Informationen enthaltend	enthält für Weitergabe nicht abtrennbare Zusatzinformationen

Quelle: [1]

Beispiel

- Biometrische Merkmale (Fingerabdruck, Gesicht)
  - stabil über die Zeit, kaum änderbar, nicht übertragbar, langfristig speicherbar (Reisepass), häufig wiederverwendet (Passkontrolle), authentisch, andere können darauf zugreifen (sofern nicht verschleiert)
  - **potentiell gefährlich für die Privatheit**

- selbstbestimmte Identitätsmerkmale (Login, Passwort)
  - anonym oder pseudonym, leicht änderbar, übertragbar an Dritte, Zugriff anderer nicht möglich, oft nur für unwichtige Teile des pers. Lebens relevant
  - **weniger gefährlich für die Privatheit**

Fallbeispiel: Verkettung von Suchanfragen in Suchmaschinen

Suchmaschinen: Google, AOL, MSN und Co.

- Suche übermittelt an Suchmaschinenbetreiber:

Kategorie	Attribute
Suchterme (technisch unvermeidlich)	Aneinanderreihung von Schlüsselworten
Browser-Kommunikation (technisch unvermeidlich)	Zeitstempel, IP-Adresse, Browser, Betriebssystem Spracheinstellungen, zuletzt besuchte Seite
Browser-Kommunikation (technisch vermeidlich)	ausgewähltes Suchergebnis, (Implementiert als Redirect)
Zusatzinformationen (vom Nutzer vermeidbar)	Cookie-Informationen, Session-ID
	über JavaScript ermittelte Daten über Mausbewegungen, Verweildauer auf der Seite, Bildschirmauflösung
	Nutzer-ID (bei persönlichem Login, z.B. Google)

Verkettung von Suchanfragen

- Über die **Browser-Kommunikation**, insbes. IP-Adresse
  - dynamische IP-Adressen: überdauern mindestens eine Such-Session
    - Kombination aus Standort des IP-Adressbereichs, Betriebssystem, Sprache, Browser etc. können als Quasi-Identifier ausreichen
  - statische IP-Adressen, z.B. Uni-Netz: länger gültig
- Über **Cookies, Session-IDs, Nutzer-Login**
  - Identifiziert einen Browser (und damit oft dessen Benutzer) über lange Zeiträume eindeutig
  - auch bei wechselnder IP-Adresse
- Über die **Suchterme**
  - z.B. Suche nach eigenem Namen, seltene Hobbies
- **Kombinationen** aus allem

Verkettung mit Zusatzdiensten

- Suchmaschinenanbieter oft Anbieter weiterer Dienste
  - Google: Youtube, Maps, Email, Verzeichnisdienste, Google Docs, Google Earth, News, Usenet
- *separate* digitale Teilidentitäten werden verkettbar
  - über IP-Adresse Nutzerbewegungen über mehrere Dienste hinweg nachvollziehbar
  - oft übergreifendes Login für viele Dienste,
    - Microsoft Passport, Windows Live ID
    - Google Authentication for Web Applications (OAuth)
- Informationen über viele Lebensbereiche
  - Arbeit, Privatleben, Hobbies, Kommunikationspartner etc.

Haben Suchanfragen Personenbezug?

- kommt ganz drauf an...
  - zur Erinnerung: *"...Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person."*
  - Heute übliche Interpretation: kein Personenbezug, wenn "*für Einzelangaben zu einer Person die Wahrscheinlichkeit, dass diese der Person zugeordnet werden können, so gering ist, dass sie nach der Lebenserfahrung oder dem Stand der Wissenschaft praktisch ausscheidet.*"

Quelle: Roßnagel, A.; Scholz, P.: Datenschutz durch Anonymität und Pseudonymität, MMR 2000

Personenbezug von Suchbegriffen

- auch hier: kommt drauf an...
  - “Katzenfutter billig” → nein;
  - “Erik Buchmann Urlaub Italien” → ja
- Personenbezug ist *abhängig von den Benutzereingaben*; nicht ohne weiteres automatisch vom Betreiber entscheidbar!
  - je umfangreicher die Suchhistorie eines Nutzers, desto wahrscheinlicher kommt eine identifizierende Kombination von Identitätsattributen zusammen *(Beispiel folgt)*

Personenbezug der Kommunikation

- Dynamische IP-Adresse
  - eher nicht personenbezogen
    - *Personenbezug erfordert Mithilfe des Kenners der Zuordnungsregel*, d.h., Internet Service Provider
- Statische IP-Adresse für festen Rechner
  - Suchmaschinenbetreiber kann Suchanfragen über lange Zeiträume einer Person zuordnen
  - mit zunehmender Zahl der Suchvorgänge steigt Wahrscheinlichkeit, dass sich der Suchende offenbart *(vgl. vorangegangene Folie)*
  - daher: oftmals personenbezogen

Anmerkung: derzeit unterschiedliche Rechtsauffassungen;
hier vorgestellt wird akutell gängige Praxis – kann sich aber ändern

Personenbezug von Zusatzinformationen

- Bildschirmauflösung, Sprache, Verweildauer auf der Seite
  - nicht als identifizierendes Merkmal geeignet
- Nutzer-Login, Session-ID, Cookie-Informationen
  - Betreiber hat Login, Session-ID bzw. Cookie selbst dem Nutzer zugeordnet
  - daher: oftmals personenbezogen, da *Suchmaschinenbetreiber gleichzeitig Kenner der Zuordnungsregel*

Aug. 2006: AOL-Datenleck (1/2)

- AOL Research veröffentlicht für Forschung 20 Mio. Suchanfragen von 650.000 Usern, gesammelt über 3 Monate
  - IDs pseudonymisiert, künstl. Schlüssel

User ID	Search Keywords	Date	Website
4417749	numb fingers	2006-03-06 18:35:02	http://irgendwas.de
...	...	...	...

- zwar ist keiner der Datensätze unmittelbar personenbezogen, aber schnell werden einzelne Identitäten und komplette Persönlichkeitsprofile offenbar

- 3 Tage später: Datenbank ist vom Netz, aber schon vielfach in Tauschbörsen kopiert, bis heute verfügbar

Aug. 2006: AOL-Datenleck (2/2)

- Nutzer 4417749: hunderte Suchanfragen in 3 Monaten
  - "dog urinates on everything": *Hundebesitzer*
  - "60 single man": *einsame ältere Frau*
  - "numb fingers": *körperliche Gebrechen*
  - "homes sold in gwinnett county": *Wohnung*
  - "xxx Arnold, yyy Arnold": *suche nach Verwandten*
  - "school supplies for Iraq children": *karitativ*
  - "best season to visit Italy": *nächster Urlaub*
- identifiziert als Thelma Arnold:
  "My goodness, its my whole personal life!"
  - Privat- und Alltagsleben, Ängste, Gebrechen
  - falsches Selbstbild durch Suchanfragen für Freunde

Quelle: http://www.nytimes.com/2006/08/09/technology/09aol.html

Wie handhaben es die Betreiber 2009?

- Google
  - nutzt Cookies, verknüpft Daten der Suchmaschine mit hauseigenen Diensten
  - Speicherdauer: nach 9 Monaten werden Suchlogs anonymisiert (was immer das heißt)
- Microsoft, AOL, Yahoo
  - vergleichbare Praxis der Datensammlung
  - Speicherdauer Yahoo, AOL: 6 Monate
  - Speicherdauer Microsoft: *“After 18 months, we will completely anonymize all Search queries [...] by irreversibly removing all cross-session identifiers [...] including the full IP address and all cookie IDs.”*
    → schützt nicht vor persönlichen Suchtermen

Zukunft: Collaborative Search Engines (CSE)

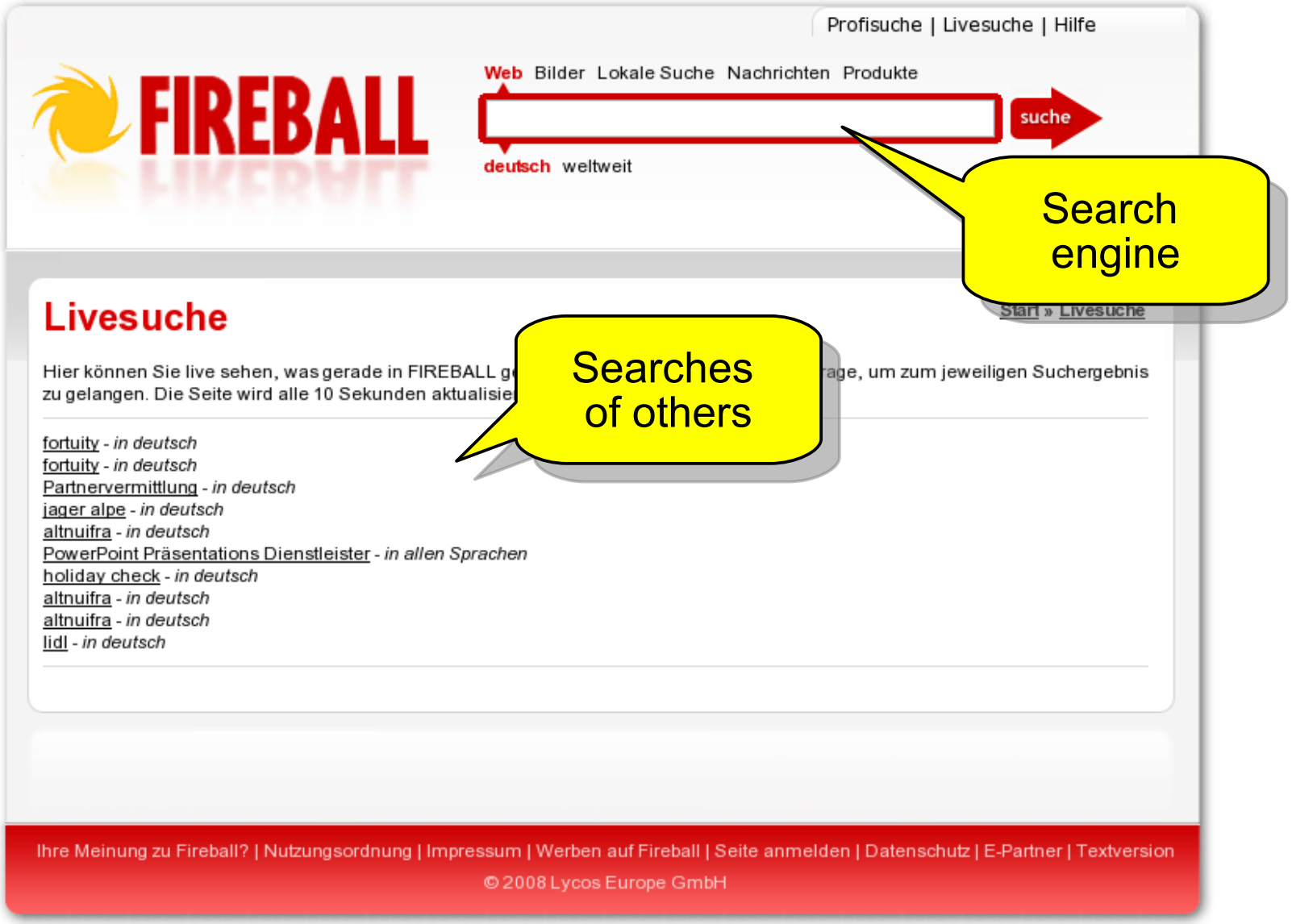
- Neuer Trend für Internet-Suche
  - effizient gemeinsam mit Freunden, Kollegen suchen
  - Unterstützen von unerfahrenen Nutzern
  - spannende Suchen und Suchergebnisse weitergeben
  - andere am Alltagsleben teilhaben lassen

Effiziente Suche

Intensive Nutzerinteraktion

Datenschutz?

Einfachstes Beispiel: Fireball Livesearch



Komplexeres Beispiel: SearchTogether

Datenschutz in aktuellen CSE-Prototypen

- **Gibts sowas?**

Microsoft SearchTogether (Beta)

April 24, 2008: SearchTogether is now available for download!
See below for details on how to install and use SearchTogether.

Privacy Policy

SearchTogether allows you to share your web search results and activity with others. When you join a SearchTogether session, your profile, search queries, results, and web site navigation, along with how you use the SearchTogether tool will be sent to Microsoft for research purposes. In addition, this information will be shared with all the people in that session, regardless of when they join. For more information, read our privacy statement.

Bedrohungen für den Datenschutz

- Jede CSE-Komponente selbst ist problematisch
  - **Suchmaschine**: Nutzerinteressen, Absichten
  - **Weitergabe von Anfragen, Links:** unkontrollierte Verbreitung persönlicher Infos
  - **Kommunikation:** private Gespräche, Kontakte
  - **Speicherung**: Anfragen, angeklickte Links, Kommunikation bleiben für lange Zeit verfügbar
- Verkettung von Informationen potenziert Gefahren
  - "**Speichere** dauerhaft, **wer** sich **wann** für **was** interessiert hat, und **mit wem** er das diskutiert hat."

Zusammenfassung

Zusammenfassung

- digitale Identitäten als Untermenge aller technisch abbildbaren Attribute mit Personenbezug
- digitale (Teil-)Identitäten sind grundsätzlich bedrohlich, so harmlos sie auch zunächst scheinen mögen
- Identitätsdiebstahl als derzeit größte Bedrohung durch den unbedachten Umgang mit pers. Informationen
- Verkettung digitaler Identitäten führt zu Informationsanreicherung und Profilbildung
- Aktuelles Fallbeispiel: Verkettung in Suchmaschinen

Literatur

[1] Verkettung Digitaler Identitäten, *Unabhängiges Landeszentrum für Datenschutz Schleswig Holstein, 2007*

[2] Thorben Burghardt, Erik Buchmann, and Klemens Böhm. *Why Do Privacy-Enhancement Mechanisms Fail, After All?,* W2Trust'08