

# Datenschutz, Big Data und das öffentliche Gut

Berlin, 6. KSWD, 21.02.14.

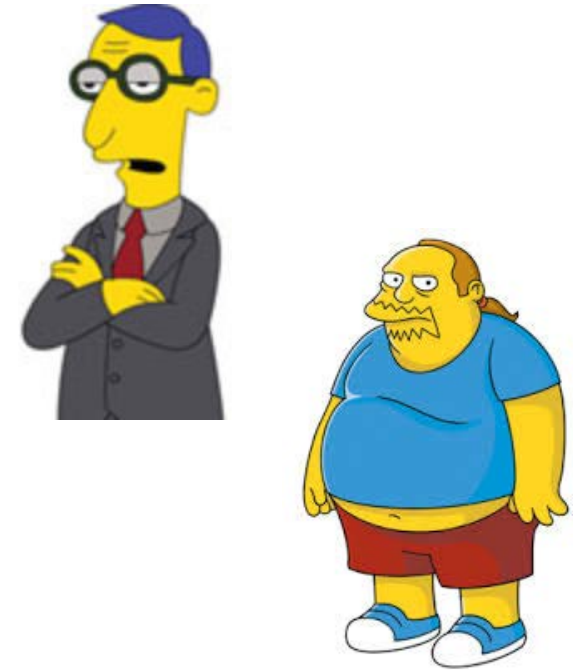
Stefan Bender (IAB)

Julia Lane (American  
Institutes for Research)

Victoria Stodden (Cornell)

Mike Holland (CUSP)

- Ich bin kein Jurist
- Ich bin auch kein Informatiker
- Vielleicht ganz gut so



Werde daher überwiegend Argumente aus Statistik und Sozial-/Wirtschaftswissenschaften verwenden.

Dank an den DiskAB von 18.2.14 (insb. Frauke Kreuter)



## **PRIVACY, BIG DATA AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT**

Julia Lane  
Victoria Stodden  
Helen Nissenbaum  
Stefan Bender

Erscheint im Herbst bei Cambridge University Press

- Wege zu identifizieren, in dem diese neuen (personen-bezogenen) Datenmengen gesammelt, integriert und analysiert werden können, um evidenzbasierte Entscheidungen bei gleichzeitiger Einhaltung des Datenschutzes zu ermöglichen:
  - Grenzerweiterungen zu legalen, statistischen und operationalen Aspekten des Datenschutzes.
  - Theoretische und praktische Grundlagen für staatliche Einrichtungen und/oder amtliche Statistik, ihre Datenzugangsregeln und Datensicherheitsprozeduren aufzubauen bzw. zu ändern.
- Im Vortrag werden sehr viel Inhalte aus dem Buch verwendet. Die Autoren sind durch kenntlich gemacht.

- Gesellschaft kann die Auswüchse der Digitalisierung nicht hinnehmen.
- Ohne Einwilligung der Betroffenen werden Daten gespeichert.
- Datenschutz stärken.
- Auf der anderen Seite viele Möglichkeiten für die Forschung.

- Big Data und Datenschutz
  - Ein Beispiel (aus vielen) als Basis
  - Informierte Einwilligung/Anonymisierung und Big Data (Barocas/Nissenbaum)
  - Differential Privacy (Dwork)
  - Datenzugang für die Wissenschaft zu Big Data
  - Schluss
- 
- Fokus des Vortrags: Wissenschaftssphäre



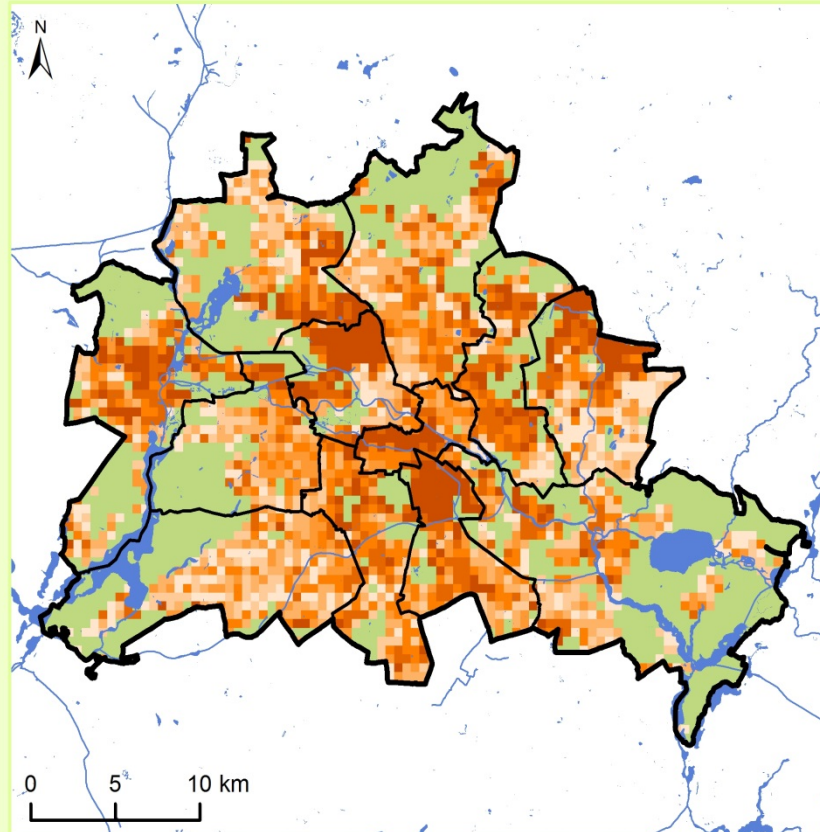


- Big Data ist ein Paradigmenwechsel.
- Trend zu immer detaillierteren und zeitnäheren Daten.
- Wechsel von einer modellgetriebenen zu einer daten- und modellgetriebenen Herangehensweise.
- Gegensatz zwischen Small Data und Big Data.
- Quantität führt zu einer neuen Art von Qualität (vom Bild zum Film).
- „Die Daten sprechen lassen“, Sekundärverwendung!

- Den Gegensatz Small vs. Big gibt es so nicht.
- Sehr viele Daten/Anwendungen der amtlichen Statistik sind Big Data-Daten/-Anwendungen (Bundesagentur für Arbeit als ein Beispiel).
- Auch oftmals Sekundärverwendung.
- Besser: Big Data by Design vs. Organic Big Data (Rob Groves).

**Abbildung 2:**

Anteil der Niedriglohnbeschäftigten  
an allen sozialversicherungspflichtig  
Beschäftigten in  
**Berlin**



Anteil in %

- 0,00 - 0,15
- 0,16 - 0,19
- 0,20 - 0,21
- 0,22 - 0,24
- 0,25 - 0,28
- 0,29 - 0,34
- 0,35 - 0,74

- Stadtgrenze
- Bezirksgrenze
- Wasserfläche
- Grün- / Agrarfläche
- Sonstiges

Quelle:  
Georeferenzierte Daten des IAB  
Geometrien: OpenStreetMap,  
ThinkGeo, Geofabrik  
© IAB

Niedriglohn = Lohn unterschreitet zwei Drittel des mittleren Lohns in Deutschland.  
Niedriglohn = ca. 57 Euro.

- Man steht generell vor zwei großen Herausforderung, wenn man (sinnvolle) Informationen aus Daten herausziehen will:
  - a. Messen: Haben wir alle relevanten Merkmale in den Daten um die Forschungsfrage beantworten zu können? Gibt es unbeobachtete Merkmale, die unsere Analyse beeinflussen und die wir kontrollieren müssten? Liegt ein systematischer Messfehler vor?
  - b. Inferenz: Verstehen wir den Prozess, wie es zur Stichprobe kam? Sind alle für die Analyse notwendigen Einheiten vorhanden? Fehlen Einheiten systematisch? Sind Einheiten mehrmals vorhanden? Haben wir für alle Einheiten ein Messung? Wen repräsentieren diese Einheiten?
- Total Survey Framework als Rahmen.
- Surveys und amtliche Daten können diese Fragen (meist) beantworten. Big Data kann dies oftmals nicht.

- Lösungen und neue Herausforderungen: Datenverknüpfungen und Integration von Informationen
- Kooperation zwischen den „Welten“ (vgl. u.a. auch Rob Groves)

**Johannes Landvogt**

Technologischer Datenschutz, IT  
c/o Bundesbeauftragter für den  
Datenschutz und die  
Informationsfreiheit

# Datenschutz im Zeitalter von Big Data

**Aber auch:  
Thilo Weichert**

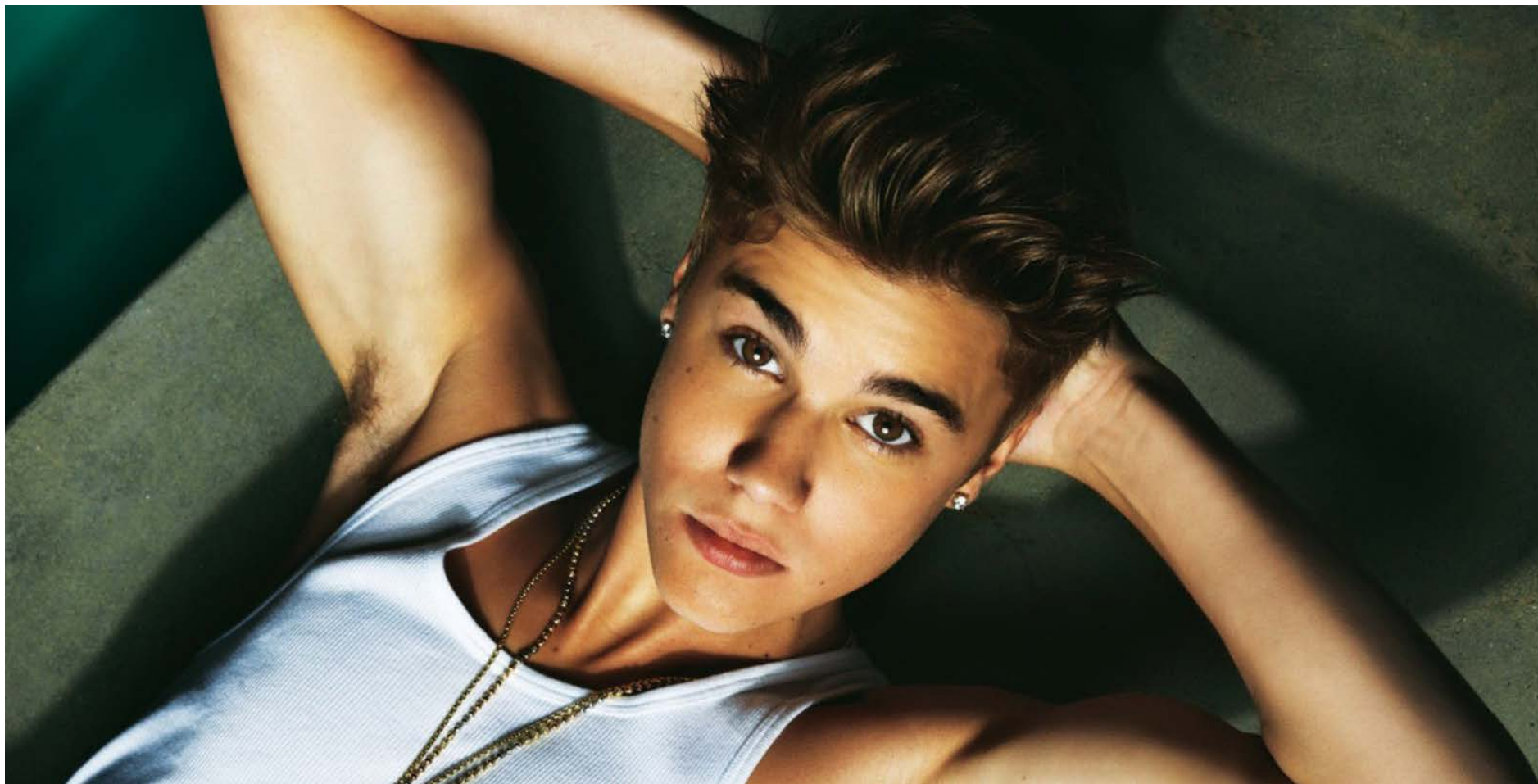
**„Big Data Days“  
im BMWi, Berlin  
12. November 2013**

- Viele Betätigungsfelder von Big Data haben einen klaren Personen- und damit Datenschutzbezug, z.B. personalisierte Werbung, Kredit-Scoring.
- Andere Big Data-Bereiche arbeiten mit anonymisierten Daten; eine Pseudonymisierung reicht (in der Regel) nicht aus, um den Personenbezug auszuschließen.

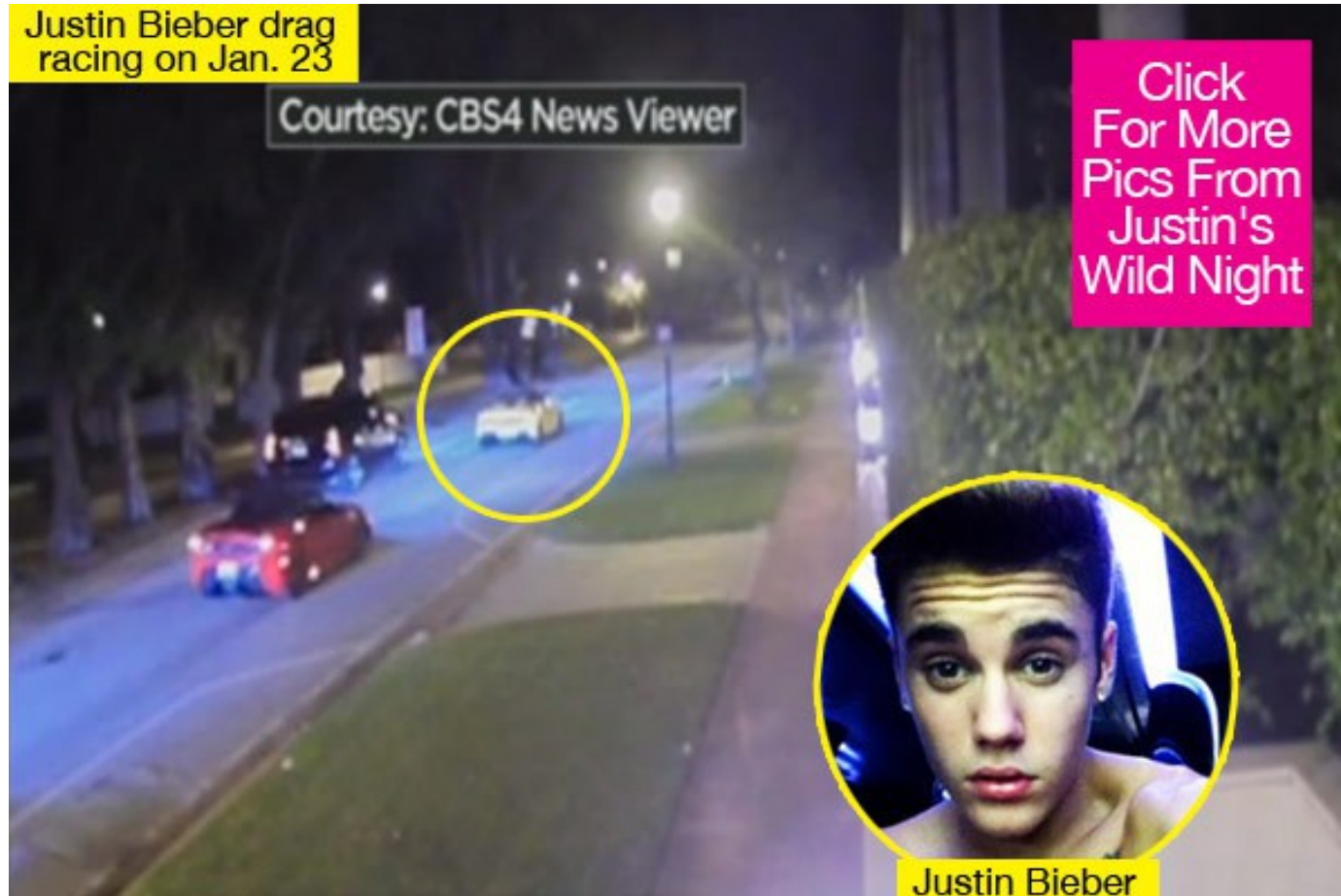
- Verbot mit Erlaubnisvorbehalt: Jede Verarbeitung personenbezogener Daten bedarf einer Rechtsgrundlage
  - Einwilligung der Betroffenen liegt häufig nicht vor.
  - Datenverarbeitung zu eigenen Zwecken (§§ 27ff. BDSG, insbes. §28)
  - Berücksichtigung der berechtigten Interesse der Betroffenen.
- Datensparsamkeit / Erforderlichkeit
- Zweckbindung
- Transparenz

## Big Data – datenschutzfreundlich?

- Beschränkung auf vorhandene Daten
  - Verwendung anonymisierter/pseudonymer Daten
  - Transparenz der Verarbeitung/Bewertung
  - Ausschluss sensibler Daten
  - Keine automatisierten Entscheidungen!
- 
- Recht auf informationelle Selbstbestimmung  
informierte Einwilligung notwendig



# Justin Bieber Drag Racing: New Video Before DUI Arrest Revealed



- Polizei: Illegales Autorennen. Biber sei verantwortungslos durch die Stadt gerast.
- Allerdings keine Radaraufnahmen.
- Im Auto war ein GPS-Ordnungssystem, was auch Fahrtdaten speichert.
- Anwälte von Biber: Rauslesen der Daten.
- Ergebnis: Biber ist nicht zu schnell gefahren (Höchstgeschwindigkeit nicht mal 40 Miles).

- Jürgen Bönninger (52. Verkehrsgerichtstage) fordert ein No-Spy-Zertifikat für Neuwagen.
- Bis zu 100 Datenverarbeitungssysteme arbeiten in einem Auto.
- Verhinderung von digitalen Abdrücken und damit der Speicherung und Abrufung von Fahrdaten sowie der – Zustandsdaten als Bewegungs- und Handlungsprofile.
- Nahezu alle Daten sind personenbezogen: vom Fahrstil bis Musikgeschmack
- Anwendungen: Versicherung, Schadensfallregulierung, Müdigkeitswarnung, ...

- Anonymität und Einwilligung sind attraktiv: Anonymisierung scheint die Daten aus der Privatsphäre zu heben.
- Aber wenn Personen nicht identifizierbar sind, können sie eventuell immer noch erreichbar sein. Sie können immer noch Gegenstand von Schlussfolgerungen oder Vorhersagen sein.

- Privatsphäre: Informationen über Personen fließen unter Einhaltung von informationellen Normen.
- Informationelle Normen bestimmen wie Informationen zwischen (Haupt-)akteuren, Informationstypen und innerhalb/über Grenzen fließen können ('transmission principles').
- Big Data hat dies Normen/ Informationsflüsse (zutiefst) verletzt („unser“ Autobeispiel)

- Der Wert einer Anonymisierung hängt nicht vom Weglassen von Namen ab, sondern ob eine potentielle Erreichbarkeit verhindert werden kann – mit oder ohne dem Zugang zu identifizierbaren Merkmalen.
- “The beauty of what we do is we don’t know who you are [...] We don’t want to know anybody’s name. We don’t want to know anything recognizable about them. All we want to do is [...] have these attributes associated with them”.  
Cindy Waxer, “Big Data Blues: The Dangers of Data Mining,” *Computerworld*, November 4, 2013

- Sehr viel kann über eine Person vorhergesagt werden, ohne etwas über die Person zu wissen. Dies gilt insbesondere im Big Data – Kontext.
- “Some years ago an engineer at Google told me why Google wasn’t collecting information linked to people’s names. ‘We don’t want the name. The name is noise.’” (Anekdote nach Hardy)



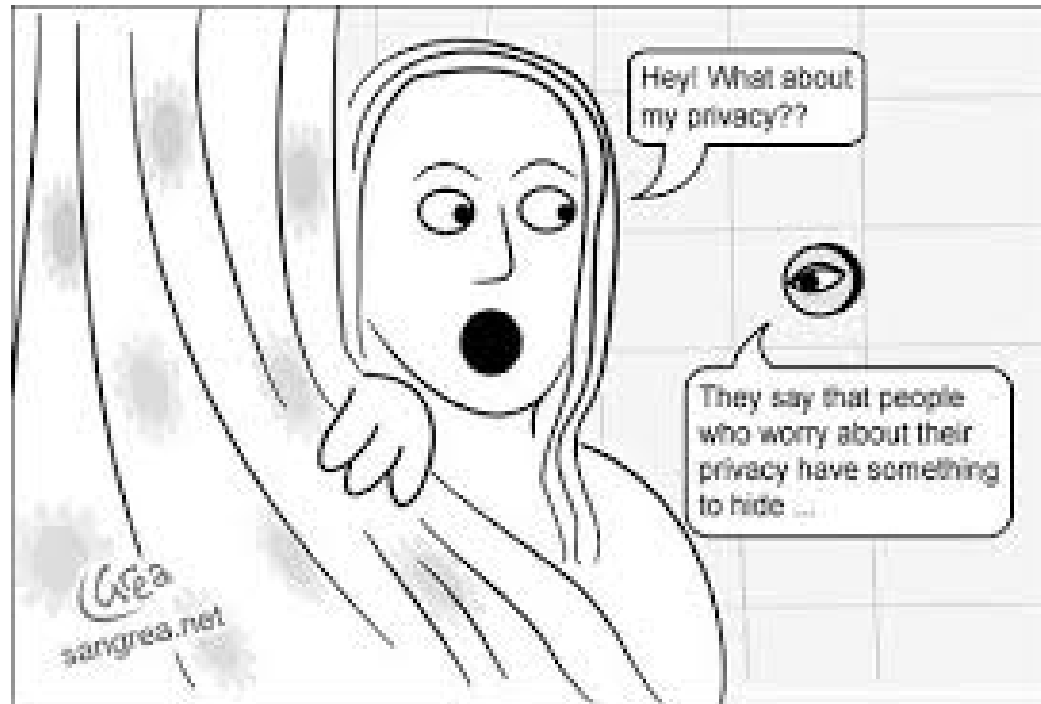
## Hat Privatsphäre einen Wert für Menschen?

- Der Wert der Privatsphäre hängt vom Kontext ab.
- Privatsphären-Paradox: Personen wollen Privatsphäre, aber nichts dafür zahlen und sind gleichzeitig bereit für sehr kleine Beträge sensitive Informationen über sich Preis zu geben.
- Personen sind nicht in der Lage die möglichen Folgen und Risiken einer (Daten-)Reidentifikation abzuschätzen (z.B. wegen bounded rationality).
- Beispiel: Datacoup will sich Nutzerdaten erkaufen. 8 \$ im Monat für Zugriff auf Kreditkartenabrechnung, Facebook und Twitter.  
(Dank an Arne Bethmann)

- Basis: Personen treffen eine autonome Entscheidung.
- Personen müssen verstehen, welche Konsequenzen ihre Einwilligung hat.
- Privatsphäre ist, die Kontrolle über die eigenen Informationen/Daten zu haben.
- Herausforderungen:
  - Transparenzparadoxon
  - Die Tyrannei der Minderheit

- Idealbild: man entscheidet sich auf einer ausreichenden Informationsbasis und weiß, was die Konsequenzen dieser Entscheidung sind.
- Informationen sollten einfach und klar sein. Das führt aber zu einem Verlust von Genauigkeit.
- Typischerweise sind Big Data-Anwendungen komplex und nicht (unbedingt) vorhersagbar.
- Einfache, geschriebene Erklärungen sind nicht ausreichend, um Entscheidungen zu Big Data-Anwendungen machen zu können.

- Die freiwillige Weitergabe von Informationen von einigen kann zur Aufdecken der gleichen Informationen von vielen führen (Notwendigkeit: Korrelation).





Oder ein wenig positiver:

„Die Informatik hat uns die Sch... eingebrockt.  
Sie soll uns da auch wieder rausholen.“  
(Freie Übersetzung von x)

- Privatsphäre und Big Data sind unverträglich.
- Dr. Ole Schröder (BMI): Big Data ist in der EU Grundverordnung nicht enthalten. Da müssen wir noch was machen. Qualität vor Geschwindigkeit.
- **Ernsthafte Diskussion über Gesetzesänderungen, die nicht nur die Datenerzeugung und –veröffentlichung im Fokus hat.**
- Strandburg: der massenhafte Erwerb, der Transfer und die Aggregation von Daten für Data Mining und vorhersagenden Analysen erzeugt zu klärende Sachverhalte, die momentan durch Gesetze nicht beantwortet werden können.
- Ohm: die momentane Gesetzgebung bezieht sich mehr auf die Sammlung und die Veröffentlichung, nicht jedoch auf die Nutzung der Daten.

- Big Data Days in Berlin SMART DATA - KONGRESS, WORKSHOP, LABOR, 12.11.2013
    - Workshop 5: Rechts- und Datensicherheit (Leitung: Dr. Alexander Duisberg, Bird & Bird München)
    - Eigentum (Prof. Thomas Hoeren, WWU Münster)
    - Datenschutz (Prof. Georg Borges, Ruhr-Universität Bochum)
    - Rechtsfragen Autonomer Systeme (Prof. Eric Hilgendorf, Universität Würzburg)
- Inhärente Internationalität der Technik provoziert einen Wettstreit der Rechtsordnungen (insbes. USA/Europa): „race to the bottom“?

- Big Data fordert eine mathematische, rigorose Theorie von Privatsphäre, eine Theorie, die auch offen ist, kumulierte Risiken der Verletzung der Privatsphäre berechnet (minimiert), da Daten analysiert, re-analysiert, mit anderen geteilt oder verknüpft werden.

- „Unsere Definition muss berücksichtigen, dass wir etwas aus den Daten lernen wollen, die wir analysieren. Es macht dabei keinen Unterschied, ob Du in den Daten bist oder nicht, weil die generalisierbaren Ergebnisse auch Dich betreffen: das ist das Herz von Differential Privacy“.
- Der Gebrauch von Daten sollte immer mit einer Veröffentlichung des Betrags an Verlust der Privatsphäre – dem Preis der Privatsphäre – einhergehen.

- Das schwierigste zu lösende Problem ist der Zugang.
- Andere Disziplinen – Ökonomie, Ethik oder Politik – können diese Daten nicht einfach nutzen, ohne dafür zu “zahlen” bzw. die Privatsphäre zu achten.

- Das ideale Szenario: Daten werden von einem vertrauenswürdigen Kurator gehalten. Die Daten bleiben geheim, nur die Ergebnisse werden publiziert (Kryptographie).
- Daten sollen nur durch einen Datenschutzmechanismus (privacy mechanism) zugänglich gemacht werden, in dem die Weitergabe von (Zwischen-)Ergebnissen minimiert wird.
- Das ist aber ein Konzept, was Forscherinnen und Forschern fremd ist, die mit Daten arbeiten.

- Datenzugang über Forschungsdatenzentren / safe center oder Remote Access.
- Karr/Reiter sehen einen freien Zugang bei stark veränderten Daten, Remote Access bei anonymisierten Daten und verification servers zur Kontrolle der Ergebnisse
- Stodden spricht sich für einen Datenzugang über „Wallet gardens“
- Und Greenwood et al. plädieren für einen „The New Deal on Data“

- Wir wollen keine Abschaffung von Bürgerrechten!!!
- Wir wollen, dass adäquate Lösungen erarbeitet werden (momentan ist die informierte Einwilligung eine Illusion)
- (Amtliche) Statistik / Sozial-/Wirtschaftswissenschaften muss aktiver werden:
  - Kein Gegensatz von small und Big Data.
  - Wir haben Wissen, was die Big Data Community braucht.
  - Big Data kostet auch uns Vertrauen (Non-Response, Qualität)
  - Wir wissen wie Datenzugang funktioniert.

- Big Data benötigt Ergebnisse aus den Sozial-/Wirtschaftswissenschaften zur Testung ihrer Ergebnisse.
- Big Data bietet immer mehr Ergebnisse an, die in das Aufgabengebiet der amtlichen Statistik fallen.
- Best of both worlds durch Kooperationen (Rob Groves, Kreuter/Peng)

<http://fdz.iab.de>

[iab.fdz@iab.de](mailto:iab.fdz@iab.de)

fon: ++49-(0)911-179-1752

fax: ++49-(0)911-179-1728

