Big Data in einer digitalisierten, datengestützten Demokratie

Ruben Bach, Frauke Kreuter

1. Einleitung

Big Data und die daraus gewonnen Informationen wirken in weite Teile moderner Demokratien hinein. Anbieter großer Internetplattformen nutzen Daten oder Datenspuren ihrer Mitglieder, um Informationsströme zu optimieren (Foster et al. 2020). Politische Entscheidungsträger nutzen Daten, um gesellschaftliche Prozesse besser zu steuern, sei das bei der Bekämpfung von Kriminalität (Lynch 2018), bei der Verbesserung der Gesundheitsversorgung (Pan et al. 2017) oder in der modernen Stadtplanung (Glaeser 2019), um nur ein paar Beispiele zu nennen. Parteien, Politikerinnen und Politiker nutzen Daten, um den Erfolg ihrer Wahlkampagnen zu maximieren (Nickerson und Rogers 2014).

Die Nutzung von Daten in diesen Kontexten ist nicht neu. Was sich verändert sind die Datentypen, die für diese Aktivitäten verwendet werden und die Art und Weise, wie sie verwendet werden. Traditionelle Datenquellen, wie ein Zensus der Bevölkerung oder andere groß aufgesetzte Bevölkerungsbefragungen, erscheinen oft zu langsam in einer Welt, in der Entscheidungen schnell getroffen werden müssen und sich soziale Gegebenheiten schnell ändern (Lane 2020). Immer häufiger werden deshalb sogenannte digitale Datenspuren verwendet.[1]

Als digitale Datenspuren werden Daten bezeichnet, die sich aus der Interaktion von Individuen mit digitalen Geräten oder Online-Informationssystemen ergeben (Howison et al. 2011, S. 769). Dazu gehören Transaktionsdaten von Zahlungssystemen, Telekommunikationsnetzen, Webseiten, Smartphone-Apps und Sensoren (Stier et al. 2020). Die Begeisterung für digitale Datenspuren rührt vor allem von der Feinkörnigkeit der Daten her, die es ermöglichen, individuelle und soziale Verhaltensweisen und Verhaltensänderungen in hoher Frequenz und in Echtzeit zu beobachten. Darüber hinaus handelt es sich um nicht-teilnehmende Messungen, d.h. die

1 Einen Überblick über digitale Datenspuren liefern Keusch und Kreuter (2021), wir verwenden in diesem Beitrag einige der dort präsentierten Materialien.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Datenerhebung erfolgt, ohne dass die beobachtete Person aktiv dazu etwas beitragen muss.[2]

In diesem Beitrag zeichnen wir nach, wo digitale Datenspuren entstehen, wenn sich Nutzerinnen und Nutzer im Internet und auf sozialen Medien bewegen und wie diese verarbeitet und genutzt werden, um Inhalte gezielt und personalisiert zu verbreiten und zu bewerben. Ein grundlegendes Verständnis der Funktionsweise und der Prozesse um die Erhebung und Nutzung von digitalen Datenspuren ist notwendig, um aktuelle politische und soziale Entwicklungen etwa zur Personalisierung von Inhalten verschiedener Art nachvollziehen und kritisch beurteilen zu können. Als praktisches Beispiel der Nutzung digitaler Verhaltensspuren wollen wir dabei einen Blick auf die gezielte Ansprache von Bürgerinnen und Bürgern mittels politischem Microtargeting werfen. Die hier diskutierten Datenquellen können selbstverständlich auch in anderen Kontexten verwendet werden.

2. *Politisches Microtargeting*

Das wohl bekannteste Beispiel der Nutzung digitaler Verhaltensdaten ist für Werbezwecke mittels *Microtargeting*, d.h. der zielgruppenbasierten Ansprache etwa auf Webseiten oder auf Social Media Plattformen wie Facebook, Instagram, Twitter und TikTok.[3] Ziele hierbei sind zum Beispiel, Werbung passgenau denjenigen Personen oder Gruppen von Personen auszuspielen, für die ein Produkt entworfen wurde, bei denen eine hohe Kaufbereitschaft vermutet wird, die eine große Reichweite besitzen, um ein Produkt in ihren Netzwerken weiter zu verbreiten, oder bei denen der größte Absatz erwartet wird. Zunehmend etabliert sich Microtargeting aber auch im politischen Raum. Politisches Microtargeting bezeichnet die Segmentierung von Personen in immer feiner definierte Gruppen anhand von Interessen, Präferenzen und Verhaltensweisen, die beispielsweise aus den digitalen Verhaltensspuren der Individuen abgeleitet werden (Kruschinski und Haller 2017). Diese Gruppen können dann mit speziell auf sie zugeschnittenen Inhalten gezielt angesprochen werden.

Leitgedanke des politischen Microtargetings ist zum einen, dass Ressourcen wie Wahlkampfmittel effizient eingesetzt werden, etwa indem politische Werbung verstärkt an die Personen ausgespielt wird, bei denen

2 Von informierter Einwilligung einmal abgesehen, dazu kommen wir später.
3 Siehe hierzu auch den Beitrag von Kelber und Leopold in diesem Sammelband.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

noch keine gefestigte Wahlabsicht vermutet wird (Nickerson und Rogers 2014; Kruschinski und Haller 2017). Zum anderen können Inhalte effektiv eingesetzt werden, das heißt, dass diese an die Zusammensetzung der Zielgruppe angepasst werden, in der Hoffnung so eine größere Wirkung zu erzielen als mit allgemein gehaltenen Inhalten. Beispielsweise könnten Zielgruppen, die vor allem junge Familien beinhalten, mit Inhalten zu Familien- und Bildungspolitik beworben werden, während Zielgruppen, die primär aus Rentnerinnen und Rentnern bestehen, verstärkt mit Inhalten zum Ausbau der Rentenversorgung angesprochen werden. Die Hoffnung hier ist, dass speziell auf einzelne Individuen oder Gruppen von Individuen mit ähnlichen Eigenschaften zugeschnittene Inhalte (etwa Wahlwerbung für die eigene Partei) deutlich effektiver sind als allgemeine Maßnahmen (Nickerson und Rogers 2014). Darauf basierend, entwickeln Datenanalystinnen und Datenanalysten mithilfe digitaler Verhaltensdaten und neuer Analyseverfahren immer detailliertere Werkzeuge zur Segmentierung von Individuen in einzelne Gruppen.

Insgesamt ist die gezielte Ansprache von (potentiellen) Wählerinnen und Wählern in Deutschland zwar noch weit weniger verbreitet als etwa in den USA. Allerdings lässt sich auch für Deutschland beobachten, dass insbesondere der Einsatz von digitalen Verhaltensspuren stetig zunimmt (Jungherr 2016). Insbesondere Social Media Plattformen wie Instagram, TikTok, Facebook und andere stellen hierfür ideale Bedingungen bereit, da ihre algorithmen- und datengetriebenen Businessmodelle auf die zielgenaue und personalisierte Ansprache ihrer Nutzerinnen und Nutzer ausgerichtet sind (Kruschinski und Bene 2021).[4]

Zu den führenden Plattformen, auf denen in Deutschland (politische) Werbung im Internet geschaltet wird, gehören YouTube, Facebook und Instagram (Kemp 2021). Die Nutzungshäufigkeit dieser Plattformen unterscheidet sich zwischen einzelnen Bevölkerungsgruppen deutlich. Das heißt auch, dass auf verschiedenen Plattformen verschiedene Bevölkerungsgruppen erreicht werden. Zudem verschiebt sich die Popularität der Plattformen regelmäßig mit der Entwicklung neuer Technologien (Beisch und Schäfer 2020). So spielen etwa insbesondere für jüngere Erwachsene Plattformen wie TikTok, Snapchat und Twitch eine deutlich größere Rolle als für ältere Generationen. Ebenso verschieben sich die selbst auferlegten Regeln einzelner Plattformen zur Schaltung politischer Werbung. Beispielsweise ist derzeit (2021) Werbung zu politischen Zwecken auf Twitter nicht

4 Siehe hierzu auch den Beitrag von Djeffal in diesem Sammelband.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

mehr erlaubt (Twitter ohne Datum)[5]. Google hingegen erlaubt politische Werbung als kontextuelle Werbung, d.h. zum Beispiel im Zusammenhang mit thematisch ähnlichen Videos auf YouTube. Ebenso möglich ist das Targeting von Personengruppen aufgrund von Alter, Geschlecht und Region mit politischer Werbung bei Google (Google 2019). Ohne größere Einschränkungen hinsichtlich der Targetingmerkmale ist politisches Microtargeting derzeit bei Facebook[6] möglich, wobei sich dies auch in naher Zukunft ändern soll (Bovermann 2021). Aufgrund der stetigen Veränderungen erläutern wir die Nutzung von digitalen Verhaltensdaten zum Zweck von Microtargeting deshalb weitestgehend unabhängig von einer bestimmten Plattform und bringen nur gelegentlich Beispiele, die sich auf einzelne Plattformen beziehen. Die Sammlung und Auswertung von digitalen Verhaltensdaten für zielgruppengenaue Werbe- und Personalisierungszwecke ist generell weit verbreitet und kann im Prinzip auf jeder Webseite, nicht nur in sozialen Medien, vorgenommen werden.

3. Daten

Die für die Entwicklung und Anwendung der für Algorithmen notwendigen digitalen Verhaltensdaten stammen in der Regel entweder von den Plattformen selbst (*Plattform-Online-Daten*), oder werden über Trackingnetzwerke erhoben (*Off-Plattform-Online-Daten*). Mitunter werden auch Offline-Daten hinzugezogen, die wir hier nur am Rande streifen.

3.1. Plattform-Online-Daten

Unter Plattform-Online-Daten verstehen wir hier alle Daten, die Nutzerinnen und Nutzer in ihrer Interaktion mit der Plattform erzeugen oder angeben. Dazu gehören etwa soziodemographische Informationen, die bei einer Registrierung angegeben werden. Verpflichtend ist bei vielen Plattformen eine Altersangabe, etwa um Volljährigkeit festzustellen oder um Kinder von der Nutzung auszuschließen. Freiwillige Angaben umfassen je nach Ausrichtung der Plattform z.B. Bildungsabschlüsse, Beziehungsstatus,

5 Das bedeutet allerdings nicht, dass Politikerinnen und Politiker und ihre Parteien nicht mit Accounts auf Twitter vertreten sein können.

6 Wir sprechen hier von der Plattform Facebook als ein Beispiel und nicht über alle Angebote des jetzt in Meta umbenannten Unternehmens.

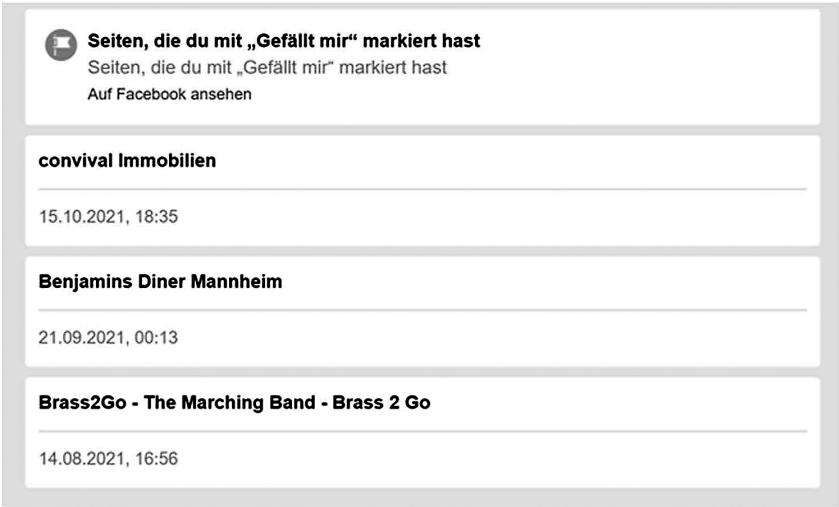
https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Geschlecht oder Informationen zum Arbeitgeber. Der Füllgrad dieser Variablen, d.h., die Anzahl der Personen, die diese Merkmale angeben, ist typischerweise eher gering (Salganik 2018, S. 24). Auch geografische Merkmale, die mit Erlaubnis der Nutzerinnen und Nutzer zu Fotos, Tweets oder Posts hinzugefügt werden, fehlen häufig. Rieder und Kühne (2018, S. 427) sprechen in ihrer Literaturübersicht beispielsweise von etwa 20% geo-getaggter Fotos auf Instagram, 10% geo-getaggten Tweets auf Twitter und etwa 10% mit Ortsangaben markierten Fotos auf Facebook.

Eher verfügbar, weil sie direkt aus der Interaktion mit der Plattform entstehen, sind Informationen zu Interaktionen mit anderen Nutzerinnen und Nutzern und deren Inhalten. Hierzu gehören Likes, Retweets, das Teilen von eigenen Inhalten, besuchte Events und Gruppenmitgliedschaften, sowie Informationen über die mit einem "Gefällt mir" versehenen oder geteilten Inhalte selbst, aber auch Interaktionen mit Unternehmensseiten und deren Produkten wie etwa Bewertungen. Auch Datum und Uhrzeit des Logins auf die Plattform können gespeichert werden. Außerdem können Informationen zum Besuch einer Seite (innerhalb von z.B. Facebook) gemeinsam mit Datum und Uhrzeit abgespeichert und über diese Merkmale mit anderen Aktivitäten zum gleichen oder ähnlichen Zeitpunkten verlinkt werden. Wie Abbildung 1 zeigt, lassen sich aus den reinen Besuchs- und Aktivitätsdaten zunächst nur sehr wenig Informationen direkt ableiten.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Abb 1. Beispieldaten von Facebook-Seiten, die mit "Gefällt mir" markiert wurden



Weitere wichtige Informationen lassen sich beispielsweise über den *user agent string* auslesen, den Nutzerinnen und Nutzer beim Aufrufen einer Webseite an den Server, der die Webseite hostet, übermitteln. Das können etwa Informationen über den genutzten Browser, das Betriebssystem, den Hersteller des Computers, den Typ des digitalen Endgeräts (PC, Smartphone, Tablet) und weitere im Browser installierte Software sein (siehe Abb. 2). Auch die IP-Adresse eines Endgeräts wird beim Besuch einer Webseite übermittelt. Über diese lassen sich Rückschlüsse auf die geografische Region ziehen, an dem sich ein Gerät und somit die das Gerät nutzende Person aufhalten. Informationen zum Betriebssystem, der Bildschirmgröße etc. werden genutzt, um zu steuern, wie die Informationen auf dem Browser angezeigt werden, z.B. ob eine für mobile Endgeräte freundliche Version der Webseite angezeigt werden muss. Forscherinnen und Forscher oder Organisationen, die Werbung schalten wollen, können Informationen über das Betriebssystem mitunter als Proxy für den sozio-ökonomischen Status der Nutzerinnen und Nutzer verwenden.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

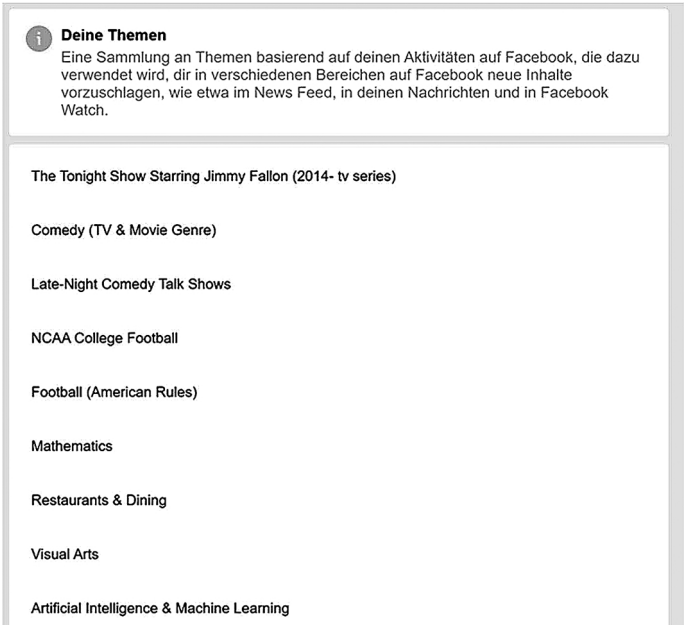
Abb 2. Beispiel eines User-Agent String

```
Your user agent: Mozilla/5.0 (iPhone; CPU iPhone OS
14_8 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like
Gecko) Version/14.1.2 Mobile/15E148 Safari/604.1
Other HTTP headers
Accept:
text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=
Accept-Encoding: gzip, deflate, br
Accept-Language: en-us
Host: duckduckgo.com
User-Agent: Mozilla/5.0 (iPhone; CPU iPhone OS 14_8
like Mac OS X) AppleWebKit/605.1.15 (KHTML, like
Gecko) Version/14.1.2 Mobile/15E148 Safari/604.1
```

Durch die Häufung von bestimmten besuchten Seiten oder die Korrelation der Häufigkeit mehrerer besuchter Seiten lassen sich thematische Präferenzen zusammenfassen. Besucht z.B. jemand innerhalb von Facebook Seiten von Restaurants, kann Dining als interessierendes Thema abgelegt werden (Abb. 3). Aufwendigere statistische Verfahren erlauben die Bildung von sogenannten Clustern ähnlicher Themen. Hierzu werden Seiten anhand möglichst vieler Merkmale kodiert (Kleidung, Art der Kleidung, Preis der Kleidung, Designerkleidung etc.) und anhand statistisch geschätzter Ähnlichkeiten andere Seiten oder Themen gefunden, die interessant sein könnten.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Abb 3. Beispielliste für eine Person auf Facebook ermittelte Themen



Ähnlich zu der Gruppierung von Seiten lassen sich auch Nutzer und Nutzerinnen anhand beobachteter Daten mittels Clustering, einem Verfahren des sogenannten *unsupervised machine learnings* in Gruppen einteilen. Ziel hierbei ist es, Cluster von Personen zu bilden, sodass die Individuen innerhalb eines Clusters möglichst ähnlich sind, zwischen den Clustern jedoch möglichst verschieden. Als Ähnlichkeitsmerkmale könnten alle oben genannten Informationen wie Alter, Einkommen, Weltanschauung aber auch Social Media Aktivität, Interessen und Präferenzen genutzt werden. Die Inhalte können dann spezifisch auf die Personen in einem Cluster zugeschnitten werden. Ein bekanntes Beispiel aus der Markt- und Sozialforschung für unsupervised machine learning sind die Sinus Milieus. Diese fassen Menschen mit ähnlichen Wertvorstellungen und einer vergleichbaren sozialen Lage in zehn Cluster, sogenannte Milieus, zusammen (Flaig und Barth 2018). Die resultierenden Milieus sind sowohl durch soziale Lage (Unterschicht bis Oberschicht) als auch Grundorientierung (Tradition bis Neuorientierung) definiert.

Fehlen für manche Nutzerinnen und Nutzer Informationen, die bei anderen vorhanden sind, so lassen diese sich mit Techniken wie dem *supervised machine learning* ergänzen oder imputieren. Hierbei wird z.B.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

geschätzt, wie wahrscheinlich eine Person, die ihr Alter, Geschlecht, ihre Bildung und ihren Beziehungsstatus angegeben hat, eine bestimmte Partei präferiert (fehlende Information), basierend auf den politischen Präferenzen anderer, ihr ähnlicher Personen, die neben ihrer Parteipräferenz demographische Angaben hinterlegt haben.

Wenn Plattform-Online-Daten mit anderen Informationen verknüpft werden, z.B. mit Antworten zu einem Persönlichkeitstest, dann kann auf diese Weise auch gelernt werden, welche Merkmale und Merkmalskombinationen welchen Persönlichkeitstyp vorhersagt. Ein Beispiel aus der akademischen Forschung für Vorhersagen solcher unbeobachteter Merkmale (wie Persönlichkeiten) anhand beobachteter Merkmale (z.B. Likes) ist die Studie von Kosinski et al. (2013). Die Autoren konnten zeigen, dass die in der psychologischen Persönlichkeitsforschung weit verbreiteten *Big Five Persönlichkeitsmerkmale* (Offenheit für Erfahrungen, Gewissenhaftigkeit, Extraversion, Verträglichkeit und Neurotizismus) mittels statistischer Vorhersagemethoden aus nur wenigen Likes, die auf Facebook abgegeben wurden, vorhergesagt werden können. Die Erkenntnisse aus dieser Studie wurden später unter anderem von der Datenanalysefirma Cambridge Analytica in ihrem datengetriebenen Modell der gezielten Ansprache von Wählerinnen und Wählern aufgegriffen.

Auch sensible Informationen wie sexuelle Orientierung, politische Ansichten oder Gesundheitsinformationen lassen sich auf diese Weise potentiell aus beobachteten Informationen abschätzen (Cabañas et al. 2018). Solange für einen Teil der Nutzerinnen und Nutzer sowohl die Merkmale, die zur Vorhersage genutzt werden (die *predictors* oder *inputs*) als auch die Merkmale, die vorhergesagt werden sollen (die *outcomes* oder *outputs*), beobachtet werden, lässt sich ein statistisches Modell für jegliche beobachteten Outcomes trainieren. Wie *genau* diese Vorhersagen sind, hängt jedoch von einer Reihe von Faktoren ab.

In anderen Worten heißt das, nur weil sich ein Vorhersagemodell trainieren lässt, bedeutet dies noch lange nicht, dass die Vorhersagen auch zutreffen. Kosinski et al. (2013) berichten beispielsweise, dass das Geschlecht, Ethnizität und ob ein Mann homosexuell ist, vereinfacht gesprochen, anhand von Social Media Likes in den ihnen zur Verfügung stehenden Daten in etwa 90% der Fälle korrekt vorhergesagt werden könne. Merkmale wie Konsum von Alkohol oder Drogen, Homosexualität von Frauen und politische Einstellungen lassen sich mit ihren Modellen jedoch deutlich schlechter vorhersagen. Das heißt, es lässt sich zwar anhand der beobachteten statistischen Zusammenhänge eine Vorhersage treffen, diese trifft aber möglicherweise in vielen Fällen nicht zu. Eine Studie des Pew Research Centers in den USA kam zum Beispiel zu dem Schluss, dass die von Face-

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

book für US-Nutzerinnen und Nutzer abgeschätzte politische "Affinität" für mehr als ein Viertel der Personen *nicht* zutrifft (Hitlin und Raine 2019). Um dies zu zeigen, wurden Teilnehmerinnen und Teilnehmer der Pew-Studie gebeten, die ihnen von Facebook zugeschriebene politische Affinität in den Einstellungen ihres Accounts abzulesen und anzugeben, ob diese ihre tatsächlichen politischen Ansichten trifft oder nicht. Da die statistischen Modelle und Algorithmen als Betriebsgeheimnis der Öffentlichkeit verborgen bleiben, liegen insgesamt wenige Erkenntnisse vor, wie präzise die Algorithmen Vorhersagen treffen können. Die Frage etwa, ob die Vorhersagemodelle der ehemaligen Datenanalysefirma Cambridge Analytica besonders präzise waren, wurde von verschiedenen Seiten angezweifelt (siehe z.B. Chen und Potenza 2018). Wie wichtig die Präzision der Vorhersage ist, hängt aber davon ab, was im Nachgang mit den gewonnenen Informationen geschieht.

Insgesamt bleibt festzuhalten, dass sich unter Verwendung einer Vielzahl an Daten, die zum Beispiel auf Social Media und auf anderen Onlineplattformen anfallen, auch nicht direkt beobachtete Informationen abschätzen lassen. Oft ist für Außenstehende jedoch nicht klar, wie zutreffend diese geschätzten Informationen sind. Der Fundus an Daten, die für die gezielte Ansprache einzelner Personen oder Gruppen von Personen genutzt werden können, kann durch die Kombination von direkt beobachteten und mittels statistischer Verfahren abgeschätzter Informationen gerade bei Plattform-Online-Daten schnell groß werden. Einen Einblick in die Informationen, die zum Beispiel Facebook über seine Nutzerinnen und Nutzer bereithält, lässt sich unter www.facebook.com/dyi gewinnen.

3.2 Off-Plattform-Online-Daten

Besonders ergiebig und nützlich werden Daten aus digitalen Verhaltensquellen, wenn Plattform-Online-Daten mit weiteren Daten aus anderen (Online-)Quellen verknüpft werden können. Wir wollen letztere hier als *Off-Plattform-Online-Daten* bezeichnen. Ein Kerninstrument der Off-Plattform-Online-Daten sind Cookies, die sich zur Sammlung von digitalen Verhaltensdaten über einzelne Plattformen und Webseiten hinaus eignen. Weitere Techniken des Trackings von Nutzerinnen und Nutzern sind z.B. Browser und Canvas Fingerprinting (Libert 2015) und die Nutzung von Advertising Identifiers, insbesondere auf mobilen Geräten wie Smartphones und Tablets (Kollnig et al. 2021). Fingerprinting-Techniken arbeiten durch das Wiedererkennen von Nutzerinnen und Nutzern anhand von (nahezu einzigartigen) Kombinationen etwa aus Gerät (Marke, Her-

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

steller, Modell und weitere Merkmale), dem genutzten Browser und der auf einem Gerät installierten Schriftarten. Advertising Identifiers (Ad-IDs) sind Identifikationsnummern, die auf Android- und iOS-Geräten genutzt werden, um Nutzerinnen und Nutzer beispielsweise über Apps hinweg verfolgen zu können und Werbepartnern die Möglichkeit zu geben, personalisierte Werbung zu schalten. Aufgrund der Omnipräsenz und Bekanntheit von Cookies fokussieren wir uns diesem Beitrag auf diese.

Cookies sind kleine Textdateien, die bei Besuchen von Webseiten von den Betreibern der Webseiten auf digitalen Endgeräten wie Computern und Smartphones der Besucherinnen und Besucher abgelegt werden (Gomer et al. 2013; Urban et al. 2018). Cookies erlauben die Re-Identifikation von Nutzerinnen und Nutzern bei wiederholten Website-Besuchen, aber auch das Sammeln von Nutzeraktivitäten über verschiedene Webseiten hinweg. Sogenannte *first-party cookies* werden genutzt, um das Browsen auf Webseiten angenehmer zu gestalten. Sie werden von der besuchten Website (der *first-party*) gesetzt und ermöglichen beispielsweise, dass Nutzerinnen und Nutzer Spracheinstellungen nicht bei jedem Websitebesuch neu konfigurieren müssen, oder dass sie bei einem erneuten Aufruf einer Website automatisch in ihren Account eingeloggt sind. *Third-party cookies* hingegen werden zwar durch die first-party gesetzt, laden jedoch Informationen, die außerhalb der besuchten Website, also bei einer third-party, liegen. Durch diese externe Referenz können Informationen über den Besuch der first-party Website mit einer third-party Webseite ausgetauscht werden.

Die third-parties sind dabei oft Werbeunternehmen, die Cookies auf sehr vielen Webseiten im Internet als third-party cookies einbinden lassen. Besucht eine Nutzerin oder ein Nutzer nun beispielsweise eine zweite Website, auf der das gleiche third-party cookie eingebunden ist, so ist die Information, dass ein und dieselbe Person beide Websites besucht hat, für das third-party Werbeunternehmen ersichtlich. Sind die Cookies einer third-party nun auf sehr vielen Webseiten eingebunden, lassen sich detaillierte Informationen über die Onlineaktivitäten einzelner Personen sammeln. Da jedoch nicht immer cookies von allen third-parties auf einer Webseite eingebunden sind, tauschen gelegentlich third-parties die in Cookies genutzten Informationen zur Wiedererkennung einzelner Nutzerinnen und Nutzer auch untereinander aus (Urban et al. 2018). So lassen sich auch dann Onlineaktivitäten für eine third-party beobachten, wenn diese selbst kein entsprechendes Cookie eingebunden hat, aber eine andere. Ist z.B. ein Cookie einer anderen third-party eingebunden und die beiden third-parties tauschen die von ihnen genutzten Informationen zur Identifikation einer Person untereinander aus, so können verschiedene

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

third-parties durch *Cookie Synchronisierung* die beobachteten Onlineaktivitäten untereinander teilen und vervollständigen. So lässt sich sicherstellen, dass die beobachteten Daten ein möglichst vollständiges Bild der Onlineaktivitäten einer Person zeichnen, auch wenn ihre eigenen Cookies nicht zwingend auf jeder Website eingebunden sind.

Schätzungen bezüglich der Verbreitung von Cookies zum Zweck der Sammlung von Onlineaktivitäten von Nutzerinnen und Nutzern gehen davon aus, dass bis zu 99% der populärsten Webseiten im Internet potentielle third-party cookies einsetzen (Kontaxis und Chew 2015; Libert 2015). Anhand der durch Cookies gesammelten Daten können Trackingunternehmen schätzungsweise bis zu 73% der Internetaktivitäten von durchschnittlichen Nutzerinnen und Nutzern beobachten (Englehardt et al. 2015; Yu et al. 2016). Die Sammlung von Nutzerverhalten durch Cookies wird dabei dominiert von einigen wenigen Unternehmen, allen voran Alphabet, der Mutterfirma von Google, sowie Meta/Facebook (Binns et al. 2018; Brandtzaeg et al. 2019; Englehardt und Narayanan 2016). Diese Unternehmen sind zugleich auch diejenigen, die einen enormen Datenfundus aus Nutzungsaktivitäten innerhalb der eigenen Plattformen generieren können, wie wir weiter oben beschrieben haben.

Sind Nutzerinnen und Nutzer in ihren Account eingeloggt oder haben nach dem Ausloggen aus ihrem Account die entsprechenden Cookies nicht gelöscht, so lassen sich die Daten aus Off-Plattform-Aktivitäten und Online-Plattform-Aktivitäten, also das Besuchen von Websites außerhalb der eigenen Plattform, leicht verknüpfen. Durch diese Kombination der Daten entstehen für Plattformen große Datenpools, die, insbesondere in Kombination mit (un)supervised machine learning Algorithmen gut für die zielgenaue und personalisierte Ansprache von Individuen genutzt werden können.

Die Fülle von Unternehmen, die third-party Cookies setzen und über Webseiten hinweg Daten sammeln geht weit über die genannten Unternehmen hinaus und das Verfolgen von Nutzeraktivitäten ist nahezu ubiquitär (Christl 2017). Normativ kritisch kann dies werden, wenn die Datensammlung in Kontexten passiert, in denen dies nicht erwartet wird. Die Philosophin Helen Nissenbaum hat auf diese Problem in dem von ihr konzipierten Framework der *Contextual Integrity* hingewiesen (Nissenbaum 2019). Zur Veranschaulichung ziehen wir die Plattform ResearchGate, eine europäischen Plattform zur Netzwerkbildung von Wissenschaftlerinnen und Wissenschaftlern, heran. Mit Stand März 2020 hat eine Nutzerin von ResearchGate, die der Voreinstellung zum Setzen von Cookies zustimmt, mit einem Schlag dem Setzen von 500 third-party cookies zugestimmt. Zudem wird durch die Zustimmung zum Setzen von Google

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Cookies rund weiteren 1500 Firmen, den Technology-Partnern von Google, Zustimmung zur Nutzung ihrer so gewonnenen Daten erteilt.

Abbildung 4: Auszug der ersten 100 third-party Cookiebetreiber, deren Cookies auf der Plattform ResearchGate eingebunden sind und so digitalen Verhaltensdaten aufzeichnen. Stand März 2020. Eine volle Liste inklusive der Technologypartner von Google befindet sich unter https://github.com/rubac/cookies_RG

1020, Inc. dba Placecast and Ericsson Emodo	Adform A/S	Adssets AB	Audience Trading Platform Ltd.
1plusX AG	Adhese	AdsWizz Inc.	AudienceProject Aps
2KDirect, Inc. (dba iPromote)	adhood.com	Adtelligent Inc.	Audiens S.r.l.
33Across	Adikteev / Emoteev	AdTheorent, Inc	AuDigent
7Hops.com Inc. (ZergNet)	ADITION technologies AG	AdTiming Technology Company Limited	audio content & control GmbH
: Tappx	Adkernel LLC	ADUX	Automattic Inc.
A Million Ads Ltd	Adledge	advanced store GmbH	Avazu Inc.
A.Mob	Adloox SA	ADventori SAS	Avid Media Ltd
Accelerize Inc.	Adludio Ltd	Adverline	Avocet Systems Limited
Accorp Sp. z o.o.	ADMAN - Phaistos Networks, S.A.	ADWAYS SAS	Axel Springer Teaser Ad GmbH
Active Agent AG	ADman Interactive SL	ADYOULIKE SA	Azerion Holding B.V.
Acuityads Inc.	adMarketplace, Inc.	Aerserv LLC	Bandsintown Amplified LLC
ad6media	AdMaxim Inc.	Affectv Ltd	Bannerflow AB
Adacado Technologies Inc. (DBA Adacado)	Admedo Ltd	Affle International	Beachfront Media LLC
adality GmbH	admetrics GmbH	Alive & Kicking Global Limited	Beemray Oy
ADARA MEDIA UNLIMITED	Admixer EU GmbH	Alliance Gravity Data Media	BeeswaxIO Corporation
AdClear GmbH	Adnami Aps	Amobee, Inc.	BEINTOO SPA
AdColony, Inc.	Adobe Advertising Cloud	AntVoice	BeOp
AddApptr GmbH	Adobe Audience Manager	Apester Ltd	Better Banners A/S
AdDefend GmbH	Adprime Media Inc.	AppNexus Inc.	BidBerry SRL
AdElement Media Solutions Pvt Ltd	adrule mobile GmbH	Arcspire Limited	Bidmanagement GmbH
Adello Group AG	Adserve.zone / Artworx AS	Arkeero	Bidstack Limited
Adelphic LLC	Adsolutions BV	ARMIS SAS	BIDSWITCH GmbH
Adevinta Spain S.L.U.	AdSpirit GmbH	Arrivalist Co.	Bidtellect, Inc
Adform A/S	adsquare GmbH	ATG Ad Tech Group GmbH	BidTheatre AB

Durch die Kombination der so erhobenen Daten mit Befragungen (eines Teils) der Nutzerinnen und Nutzer einer Webseite lassen sich zum Beispiel mit den zuvor beschriebenen Methoden des supervised machine learnings auch Modelle trainieren, die die Inhalte der Befragung dann für alle Webseiten-Nutzerinnen und -Nutzer vorhersagen können. So könnte man beispielsweise einige Personen etwa in einer Onlinebefragung nach verschiedenen Merkmalen wie soziodemographischen Informationen, Parteipräferenzen und Wahlabsicht befragen und die so gewonnenen Daten mit ihren Onlineaktivitäten verknüpfen. Anhand dieser Daten ließe sich dann ein statistisches Modell trainieren, das später angewandt werden könnte, um allein anhand der z.B. aus Cookies gesammelten Onlineaktvitäten Informationen zu soziodemographischen Informationen, Parteipräferenzen und Wahlabsicht der Onlinenutzerinnen und -nutzer zu generieren. In der Praxis zeigt sich, dass die Vorhersage basierend auf Onlineaktivitäten in der Regel für einige Merkmale wie Alter, Geschlecht, Bildung, Beruf und Einkommensgruppen gut funktioniert, für andere Merkmale jedoch keine sehr genauen Vorhersagen getroffen werden können (siehe z.B. Hinds und Joinson 2018 und Kapitel 3.1).

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Unsere eigene Forschung zu digitalen Verhaltensspuren hat gezeigt, dass Merkmale wie politische Einstellungen oder Wahlverhalten sich für Internetnutzerinnen und -nutzer in Deutschland nur mit verhältnismäßig geringer Genauigkeit aus reinen Off-Plattform-Online-Daten abschätzen lassen (Bach et al. 2021). Abbildung 5 verdeutlicht dies anhand eines Vergleichs der Vorhersagegenauigkeit verschiedener supervised machine learning Modelle. Vorhergesagt wird dabei, ob eine Person die Partei *Bündnis 90/Die Grünen* bei der Bundestagswahl 2017 gewählt hat oder nicht (linke Abbildung) bzw. die Partei *Alternative für Deutschland* gewählt hat oder nicht (rechte Abbildung). Unsere Modelle sind dabei inspiriert von Datensammlungs- und -auswertungspraktiken wie man sie auch in der Praxis vorfindet. Die Boxplots fassen die Güte der Vorhersagen anhand verschiedener Prädiktorengruppen zusammen. Die Vorhersagegüte wird dabei über die ROC-AUCs[7] der Crossvalidierungssets[8] abgebildet. Vereinfacht gesagt ist die Genauigkeit der Vorhersagen dann hoch, wenn die Werte möglichst nah an den Wert eins reichen.

In der ersten Zeile ("Soz.dem.") haben wir für die Vorhersage der Wahlentscheidung nur soziodemographische Merkmale genutzt, also etwa Alter, Geschlecht und Bildung. In der zweiten Zeile ("Websites/Apps") hingegen haben wir Informationen über die von einer Person während der vier Monate vor der Wahl besuchten Websites und auf ihrem Smartphone genutzten Apps als Prädiktoren genutzt.[9] In der dritten Zeile ("Soz.dem. + Tracking allg.") haben wir sowohl soziodemographische Informationen als auch allgemeine Informationen über das Onlineverhalten der letzten Monate einer Person genutzt, etwa die durchschnittliche Länge und die am häufigsten beobachteten Wochentage und Uhrzeiten der Internetnutzung. In der vierten Zeile ("Soz.dem. + Nachrichtenkonsum") war dagegen insbesondere der Nachrichtenkonsum von Interesse, ausgehend von der Annahme, dass der Nachrichtenkonsum einer Person Aufschluss über ihre politischen Präferenz geben könnte. In der letzten Zeile schließlich ("Soz.dem. + Websites/Apps") haben wir alle in den anderen Modellen

7 Area under the Receiver Operating Curve. Würde man für jede Person eine Münze werfen, um die Wahlentscheidung zu bestimmen, so würde sich ein ROC-AUC Wert von 0,5 ergeben. Ein Vorhersagemodell, das viele richtige Vorhersagen macht (also etwa für tatsächliche AfD-Wählerinnen und -Wähler die AfD-Wahlentscheidung auch vorhersagt), weist ROC-AUCs größer 0,5 und nahe eins auf.

8 Datenpunkte, die während des Trainings unserer machine learning Modelle genutzt werden, um aus einer Fülle von möglichen Modellen das Beste auszuwählen.

9 Für Informationen zur Sammlung dieser Daten verweisen wir aus Platzgründen auf das Papier (Bach et al. 2021).

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

enthaltenen Prädiktoren, die wir aus digitalen Verhaltensspuren abgeleitet haben, mit soziodemographischen Informationen kombiniert. Wie oben erwähnt, lassen sich jedoch mit keinem der Modelle besonders genaue Vorhersagen erzielen. Das heißt, die von uns trainierten Modelle erlauben es uns nicht, aus den digitalen Verhaltensspuren die politischen Präferenzen einer Person, gemessen über ihre Wahlentscheidung, treffgenau nachzuvollziehen. Interessanterweise zeigt sich jedoch, dass die Vorhersagen für die Wahl der AfD weniger Varianz über die verschiedenen Crossvalidierungssets aufweisen. Vereinfacht gesagt können wir daraus schließen, dass es Unterschiede in der Vorhersagegüte für die einzelnen Parteien gibt. Zwar lassen sich für keine der beiden Parteien besonders genaue Vorhersagen treffen, für die AfD funktioniert es jedoch etwas besser als für die Grünen.

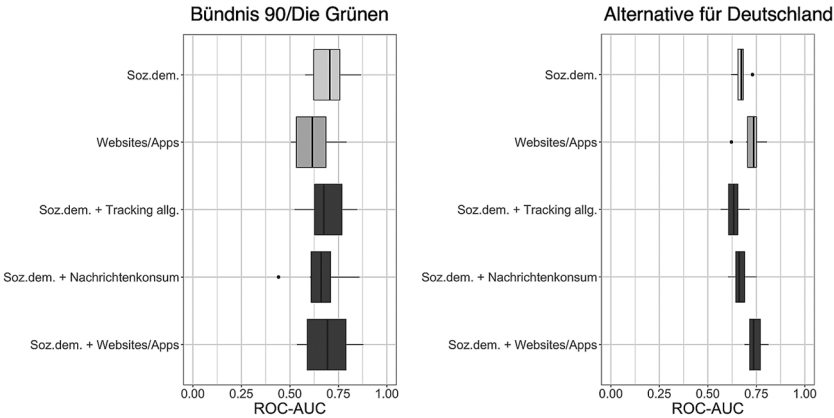
Wie oben schon angeschnitten sind die Unterschiede in der Vorhersagegüte der Modelle je nach Nutzung der Daten mehr oder weniger relevant. Kampagnen, die sich darauf konzentrieren Leute an die Wahlurnen zu bringen, können durchaus Erfolge verzeichnen und sind mitunter effektiver als solche, die versuchen Meinungen zu verändern (siehe zum Beispiel Nickerson und Rogers 2014 und Baldwin-Philippi 2019, zum Unterschied zwischen persuasion und mobilisation). Das heißt, selbst wenn die Vorhersagegüte von Modellen variiert und nicht immer zutrifft, so können sie dennoch in politischen Kampagnen hilfreich sein, um Ressourcen effektiver einzusetzen, etwa wenn für einen Teil der Personen korrekte Vorhersagen gemacht werden und diese dann etwa zur Wahl mobilisiert werden können.

Zusammenfassend lässt sich also festhalten, dass durch nahezu omnipräsentes Verfolgen von Nutzeraktivitäten im Internet große Mengen an Daten und Informationen anfallen und gesammelt werden. Diese könnten dann unter anderem zum Trainieren eines statistischen Vorhersagealgorithmus genutzt werden, der die gezielte Wählerinnen- und Wähleransprache ermöglicht. Wenn etwa für einen Teil der Nutzerinnen und Nutzer sowohl Wahlabsicht und Onlineaktivitäten beobachtet werden, könnte mit diesem Algorithmus für all die, bei denen nur Onlineaktivitäten beobachtet werden, eine Vorhersage gemacht werden, etwa ob sie unentschlossen sind, welche Partei sie wählen werden. Die so identifizierten unentschlossenen Wählerinnen und Wähler könnten dann mit gezielt auf ihre Präferenzen zugeschnittenen Inhalten angesprochen werden. Solange der Algorithmus jedoch nicht selbst trainiert wird oder offen einsehbar ist, wie gut die Identifikation der unentschlossenen Wählerinnen und Wähler ist, bleibt die Qualität der Vorhersagen offen. Die bisherige Forschung, die sich dieser Frage widmet, deutet daraufhin, dass für einige Merkmale

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

genaue Vorhersagen getroffen werden können, während sich für andere Merkmale nur unscharfe Vorhersagen machen lassen.

Abbildung 5: Die Boxplots fassen die Güte der Vorhersagen bezüglich der Wahlentscheidung einer Person für oder gegen die genannte Partei anhand verschiedener Prädiktorengruppen zusammen. Die Vorhersagegüte wird dabei über die ROC-AUCs der Crossvalidierungssets abgebildet. Vereinfacht gesagt ist die Genauigkeit der Vorhersagen dann hoch, wenn die Werte möglichst nah an den Wert eins reichen.



3.3 Offline-Daten

Abschließend wollen wir noch Offline-Daten erwähnen. Eine detaillierte Behandlung dieser Datenquellen würde den Rahmen dieses Kapitels sprengen. Wir wollen sie dennoch hier erwähnen, um aufzuzeigen, welche weiteren Daten sich mit den hier genannten verknüpft lassen und beispielsweise in den USA häufig auch im politischen Microtargeting Verwendung finden (Nickerson und Rogers 2014).

Ein gutes Beispiel sind Informationen, die über Kundenkarten oder Kundenbindungsprogramme gewonnen werden. Die Verknüpfung kann zum Beispiel über Adressen, Telefonnummern oder E-Mail-Adressen, die bei der Registrierung einer Kundenkarte oder eines Kundenkontos angegeben werden, stattfinden. Wird die gleiche Adresse, Telefonnummer oder E-Mail-Adresse bei der Nutzung von Onlineplattformen angegeben, so können die Datenquellen problemlos zusammengefügt werden. Auch

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

wenn keine eindeutigen Informationen zur Verknüpfung von Informationen aus verschiedenen Quellen vorhanden sind, lassen sich Informationen mittels Record-Linkage-Verfahren, die auf statistischen Wahrscheinlichkeiten basieren, verknüpfen (siehe z.B. Tokle und Bender 2020).

Als Lieferant von Offline-Daten werden häufig Data Broker wie das Unternehmen Acxiom (https://www.acxiom.de) genutzt. Data Broker sind Unternehmen, die auf die Sammlung von Daten über Individuen aus verschiedensten Quellen sowie das Handeln und Lizenzieren dieser Daten an Dritte spezialisiert sind. Auch wenn Data-Broker in öffentlichen Debatten um Daten und Demokratie im digitalen Zeitalter oft nicht so stark im Rampenlicht stehen wie Betreiber sozialer Medien, so sind sie doch ein integraler Bestandteil von Datenströmen in modernen digitalisierten Demokratien.

4. Zusammenfassung und Diskussion

Dieses Kapitel gibt einen kurzen Einblick in die wesentlichen Datenströme und zeigt, dass aus unstrukturierten Daten mit Hilfe statistischer oder datengetriebener Verfahren strukturierte Informationen über Individuen generiert werden können. Motiviert haben wir dabei die Nutzung der Daten im Kontext von Wahlwerbung, aber selbstverständlich können diese Daten auch zu anderen Zwecken genutzt werden.

Wie eingangs erwähnt sind digitale Datenströme, egal aus welchen Quellen sie kommen, für viele von Interesse. Auch evidenzbasierte Politik ist auf Daten angewiesen, um den Zustand einer Gesellschaft zu erfassen und Veränderungen in einer Gesellschaft zu erkennen. Deshalb ist es bei der Generierung neuer Regularien besonders wichtig, die Verwendungszwecke der Daten im Auge zu behalten und nicht Datenströme per se abzuschneiden.[10] So fließen automatisiert erhobene Datenströme etwa von Kassensystemen und aus Onlinequellen mitunter in die Berechnung von Inflationsindizes ein (Leclair et al. 2019). Spätestens seit der Coronapandemie wird zudem an zahlreichen Stellen deutlich, dass Datenströme von Plattformen oder Transaktionen Informationslücken schließen konnten, etwa bei der Vorhersage der Pandemieentwicklung selbst (Salomon et al. 2021).

Derzeit liegt die Verantwortung der Datenweitergabe bei den einzelnen Nutzerinnen und Nutzern. Die Informationsdichte über die möglichen

10 Siehe hierzu auch den Beitrag von Buchmann in diesem Sammelband.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Verwendungen ist allerdings sehr hoch und selbst solche Nutzerinnen und Nutzer, die sich der verhaltensorientierten Online-Werbung bewusst sind und maßgeschneiderte Werbung und personalisierte Suchergebnisse als hilfreich empfinden, wissen oft nicht, wie und was Unternehmen aus ihren Daten lernen können (Dolin et al. 2018; Hitlin und Raine 2019; Ur et al. 2012). Auch wenn im Kleingedruckten der Einwilligungen im Prinzip nachvollzogen werden kann, mit wem die Daten geteilt werden und für welche Analysen die Daten verwendet werden, bleibt die gesamte Daten- und Analysekette häufig doch undurchsichtig (z. B. Christl 2017). Ein Grund dafür ist, dass die Details der Algorithmen, die zur Analyse der Daten verwendet werden, oft nicht bekannt sind, wodurch eine Prüfung der Angemessenheit des Informationsflusses (Nissenbaum 2019) unmöglich wird. Ob hier mehr Transparenz hilft, oder eine Verlagerung der Verantwortung ein besseres Instrument wäre, ist eine offene Debatte. Denkbar wäre zum Beispiel, dass schadhafte Nutzung von Daten nicht nur zivilrechtlich sondern auch strafrechtlich zu verfolgen, ganz unabhängig davon woher Datenströme kommen und welcher Nutzung zugestimmt wurde.

Diese Fokussierung auf die Verwendung ist vor allem auch im Hinblick darauf sinnvoll, dass die einzelnen Nutzerinnen und Nutzer ohnehin nur begrenzt Kontrolle darüber haben, welche Vorhersagen für sie getroffen werden. Selbst wenn sich Einzelne gezielt dafür entscheiden, Informationen über sich zurückzuhalten, erlauben moderne mathematische und statistische Verfahren das Imputieren fehlender Werte, und Algorithmen können darauf trainiert werden, bestimmte Informationen vorherzusagen (Bischoff et al. 2018; Christl 2017; Lecuyer et al. 2015). Wenn also genügend andere ihre Informationen teilen, können sich einzelne nicht gegen eine Inferenz auf ihre eigenen Informationen schützen. Es ist deshalb durchaus überlegenswert, die Verantwortung stärker auf die Seite der Datennutzer zu verlagern. Nissenbaums Leitgedanken zur *Contextual Integrety* können hier ein Ansatz sein, der auf einen normgerechten Umgang mit Daten plädiert.

Eine Regulierung der Verwendung anstatt einer Regulierung der einzelnen Datentypen oder Datenströme wäre auch deshalb überlegenswert, da sich derzeit ohnehin nicht absehen lässt, welche zusätzlichen Datenströme auftauchen werden, welche Verlinkungen von Datenquellen in der Zukunft denkbar sind und welche Rechenleistung zukünftig vorhanden sein wird, um Vorhersagen zu beschleunigen oder zu verbessern. Wichtig wäre es deshalb einen Rahmen zu schaffen, der flexibel genug ist Individuen vor Schaden zu schützen ohne die positive Nutzen von Daten zu blockieren.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Literaturverzeichnis

Bach, Ruben L.; Kern, Christoph; Amaya, Ashley; Keusch, Florian; Kreuter, Frauke; Hecht, Jan; Heinemann, Jonathan (2021): Predicting Voting Behavior Using Digital Trace Data. In: *Social Science Computer Review* 39 (5), S. 862–883. DOI: 10.1177/0894439319882896.

Baldwin-Philippi, Jessica (2019): Data campaigning: between empirics and assumptions. In: *Internet Policy Review* 8 (4), S. 1–18. DOI: 10.14763/2019.4.1437.

Beisch, Natalie; Schäfer, Carmen (2020): Ergebnisse der ARD/ZDF-Onlinestudie 2020. Internetnutzung mit großer Dynamik. Medien, Kommunikation, Social Media. In: *Media Perspektiven* 9, S. 462–481.

Binns, Reuben; Lyngs, Ulrik; van Kleek, Max; Zhao, Jun; Libert, Timothy; Shadbolt, Nigel (2018): Third Party Tracking in the Mobile Ecosystem. In: Proceedings of the 10th ACM Conference on Web Science. Amsterdam, the Netherlands, 27–30 May 2018. New York: ACM, S. 23–31.

Bischoff, J.; Cygan, S.; Münkel, J.; Schindler, W. (2018): Auf Datensuche in der Welt der Datenhändler. In: *mdr*, 2018. Online verfügbar unter https://www.mdr.de/datenspuren/datenspuren-138.html, zuletzt geprüft am 12.07.2019.

Bovermann, P. (2021): Facebook dreht am Anzeigenalgorithmus. In: *SZ Online*, 10.11.2021. Online verfügbar unter https://www.sueddeutsche.de/wirtschaft/facebook-targeting-werbung-abschalten-1.5461094, zuletzt geprüft am 10.11.2021.

Brandtzaeg, P. B.; Pultier, A.; Moen, G. M. (2019): Losing control to data-hungry apps: A mixed-methods approach to mobile app privacy. In: *Social Science Computer Review* 37, S. 466–488.

Cabañas, José González; Cuevas, Ángel; Cuevas, Rubén (2018): Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In: Proceedings of the 27th USENIX Security Symposium. 27th USENIX Security Symposium (USENIX Security 18), S. 479–495. Online verfügbar unter https://www.usenix.org/conference/usenixsecurity18/presentation/cabanas.

Chen, A.; Potenza, A. (2018): Cambridge Analytica's Facebook data abuse shouldn't get credit for trump: 'I think Cambridge Analytica is a better marketing company than a targeting company.'. In: *The Verge*, 2018. Online verfügbar unter https://www.theverge.com/2018/3/20/17138854/cambridge-analytica-facebook-data-trump-campaign-psychographic-microtargeting, zuletzt geprüft am 28.10.2021.

Christl, W. (2017): Corporate surveillance in everyday life: How companies collect, combine, analyze, trade, and use personal data on billions. Cracked Labs. Vienna, Austria. Online verfügbar unter https://crackedlabs.org/en/corporate-surveillance, zuletzt geprüft am 28.10.2021.

Dolin, Claire; Weinshel, Ben; Shan, Shawn; Hahn, Chang Min; Choi, Euirim; Mazurek, Michelle L.; Ur, Blase (2018): Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA. New York, NY, USA: ACM.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Englehardt, S.; Narayanan, A. (2016): Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. Vienna, Austria, 24 - 28th October. New York: ACM, S. 1388–1401.

Englehardt, S.; Reisman, D.; Eubank, C.; Zimmermann, P.; Mayer, J.; Narayanan, A.; Felten, E. W. (2015): Cookies that give you away: The surveillance implications of web tracking. In: Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 18 - 22 May. NY: ACM, S. 289–299.

Flaig, B. B.; Barth, B. (2018): Hoher Nutzwert und vielfältige Anwendung: Entstehung und Entfaltung des Informationssystems Sinus-Milieus®. In: B. Barth, B. B. Flaig, N. Schäuble und M. Tautscher (Hg.): Praxis der Sinus-Milieus®: Springer VS, Wiesbaden, S. 3–21. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-658-19335-5_1.

Foster, Ian; Ghani, Rayid; Jarmin, Ron S.; Kreuter, Frauke; Lane, Julia (Hg.) (2020): Big Data and Social Science: A Practical Guide to Methods and Tools. London: CRC Press.

Glaeser, Edward (2019): Urban Management in the 21st Century: Ten Insights from Professor Ed Glaeser: Centre for Development and Enterprise (CDE). Online verfügbar unter https://www.africaportal.org/publications/urban-management-21st-century-ten-insights-professor-ed-glaeser/.

Gomer, Richard; Rodrigues, Eduarda Mendes; Milic-Frayling, Natasa; Schraefel, M. C. (2013): Network Analysis of Third Party Tracking: User Exposure to Tracking Cookies through Search. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Atlanta, GA, 17 - 20 November. New York: IEEE, S. 549–566.

Google (2019): An update on our political ads policy. Online verfügbar unter https://www.blog.google/technology/ads/update-our-political-ads-policy/, zuletzt geprüft am 28.10.2021.

Hinds, J.; Joinson, A. N. (2018): What demographic attributes do our digital footprints reveal? A systematic review. In: *PLoS One* 13 (11), S. 1–40.

Hitlin, Paul; Raine, Lee (2019): Facebook Algorithms and Personal Data. In: *Pew Research Center*. Online verfügbar unter https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/, zuletzt geprüft am 28.10.2021.

Howison, James; Wiggins, Andrea; Crowston, Kevin (2011): Validity Issues in the Use of Social Network Analysis with Digital Trace Data. In: *Journal of the Association for Information Systems* 12 (12), S. 768–797. DOI: 10.17705/1jais.00282.

Jungherr, A. (2016): Four Functions of Digital Tools in Election Campaigns: The German Case. In: *International Journal of Press/Politics* 3, S. 358–377.

Kemp, S. (2021): Digital 2021: Germany. Online verfügbar unter https://datareportal.com/reports/digital-2021-germany, zuletzt geprüft am 28.10.2021.

Keusch, Florian; Kreuter, Frauke (2021): Chapter 7 Digital Trace Data. In: Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg (Hg.): Handbook of Computational Social Science, Vol 1: Taylor & Francis. Online verfügbar unter https://library.oapen.org/handle/20.500.12657/51412.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Kollnig, Konrad; Shuba, Anastasia; Binns, Reuben; van Kleek, Max; Shadbolt, Nigel (2021): Are iPhones Really Better for Privacy? Comparative Study of iOS and Android Apps. In: *arXiv preprint* (arXiv:2109.13722). Online verfügbar unter https://arxiv.org/abs/2109.13722.

Kontaxis, Georgios; Chew, Monica (2015): Tracking Protection in Firefox For Privacy and Performance. In: Abigail Goldsteen, Tyrone Grandison, Mike Just, Larry Koved, Rohan Malcolm und Sean Thorpe (Hg.): Proceedings of the 9th Workshop on Web 2.0 Security and Privacy (W2SP) 2015. San Jose, CA, 21.05. Online verfügbar unter https://arxiv.org/abs/1506.04104.

Kosinski, Michal; Stillwell, David; Graepel, Thore (2013): Private traits and attributes are predictable from digital records of human behavior. In: *Proceedings of the National Academy of Sciences* 110 (15), S. 5802–5805. DOI: 10.1073/pnas.1218772110.

Kruschinski, Simon; Bene, Márton (2021): In varietate concordia?! Political parties' digital political marketing in the 2019 European Parliament election campaign. In: *European Union Politics* 23 (1), S. 43–65. DOI: 10.1177/14651165211040728.

Kruschinski, Simon; Haller, André (2017): Restrictions on data-driven political micro-targeting in Germany. In: *Internet Policy Review* 6 (4), S. 1–23. DOI: 10.14763/2017.4.780.

Lane, Julia (2020): Democratizing our data. A manifesto. Cambridge, Massachusetts: The MIT Press.

Leclair, Marie; Léonard, Isabelle; Rateau, Guillaume; Sillard, Patrick; Varlet, Gaëtan; Vernédal, Pierre (2019): Scanner Data: Advances in Methodology and New Challenges for Computing Consumer Price Indices. In: *Economie et Statistique / Economics and Statistics* 509, S. 13–29. DOI: 10.24187/ecostat.2019.509.1981.

Lecuyer, Mathias; Spahn, Riley; Spiliopolous, Yannis; Chaintreau, Augustin; Geambasu, Roxana; Hsu, Daniel (2015): Sunlight. In: Indrajit Ray, Ninghui Li und Christopher Kruegel (Hg.): Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS'15: The 22nd ACM Conference on Computer and Communications Security. Denver Colorado USA, 12.10.2015 - 16.10.2015. New York, NY, USA: ACM, S. 554–566.

Libert, T. (2015): Exposing the hidden Web: An analysis of third-party HTTP requests on 1 million websites. In: *International Journal of Communication* 9, S. 1–10. Online verfügbar unter https://arxiv.org/abs/1511.00619.

Lynch, James (2018): Not Even our Own Facts: Criminology in the Era of Big Data. In: *Criminology* 56 (3), S. 437–454. DOI: 10.1111/1745-9125.12182.

Nickerson, David W.; Rogers, Todd (2014): Political Campaigns and Big Data. In: *Journal of Economic Perspectives* 28 (2), S. 51–74. DOI: 10.1257/jep.28.2.51.

Nissenbaum, H. (2019): Contextual Integrity Up and Down the Data Food Chain. In: *Theoretical Inquiries in Law* 20 (1), S. 221–256. Online verfügbar unter https://www.degruyter.com/document/doi/10.1515/til-2019-0008/html.

Pan, Ian; Nolan, Laura B.; Brown, Rashida R.; Khan, Romana; van der Boor, Paul; Harris, Daniel G.; Ghani, Rayid (2017): Machine Learning for Social Services: A Study of Prenatal Case Management in Illinois. In: *American journal of public health* 107 (6), S. 938–944. DOI: 10.2105/AJPH.2017.303711.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Rieder, Y.; Kühne, S. (2018): Geospatial Analysis of Social Media Data - A Practical Framework and Applications. In: Stuetzer, C.M., Welker, M. und M. Egger (Hg.): Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications. Köln: Herbert van Halem Verlag (DGOF Schriftenreihe), S. 417–440.

Salganik, M. (2018): Bit By Bit. Social Research in the Digital Age. Princeton, NJ: Princenton University Press.

Salomon, Joshua A.; Reinhart, Alex; Bilinski, Alyssa; Chua, Eu Jing; La Motte-Kerr, Wichada; Rönn, Minttu M. et al. (2021): The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. In: *Proceedings of the National Academy of Sciences* 118 (51). DOI: 10.1073/pnas.2111454118.

Stier, Sebastian; Breuer, Johannes; Siegers, Pascal; Thorson, Kjerstin (2020): Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. In: *Social Science Computer Review* 38 (5), S. 503–516. DOI: 10.1177/0894439319843669.

Tokle, J.; Bender, S. (2020): Big Data and Social Science. Data Science Methods and Tools for Research and Practice. In: Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter und Julia Lane (Hg.): Big Data and Social Science: A Practical Guide to Methods and Tools. London: CRC Press.

Twitter (ohne Datum): Political content. Online verfügbar unter https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html, zuletzt geprüft am 28.10.2021.

Ur, Blase; Leon, Pedro Giovanni; Cranor, Lorrie Faith; Shay, Richard; Wang, Yang (2012): Smart, useful, scary, creepy. In: Lorrie Faith Cranor (Hg.): Proceedings of the Eighth Symposium on Usable Privacy and Security - SOUPS '12. the Eighth Symposium. Washington, D.C, 11.07.2012 - 13.07.2012. New York, New York, USA: ACM Press, S. 1.

Urban, T.; Tatang, D.; Degeling, M.; Holz, T.; Pohlmann, N. (2018): The unwanted sharing economy: An analysis of cookie syncing and user transparency under GDPR. In: *arXiv preprint* (arXiv:1811.08660), https://arxiv.org/pdf/1811.08660.

Yu, Z.; Macbeth, S.; Modi, K.; Pujol, J. M. (2016): Tracking the trackers. In: Proceedings of the 25th international conference on World Wide Web. Montreal, Cananda, 11 - 15 April. New York, NY: ACM, S. 121–132.

https://doi.org/10.5771/9783748932741-127 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY