DIW Wochenbericht 27 2023

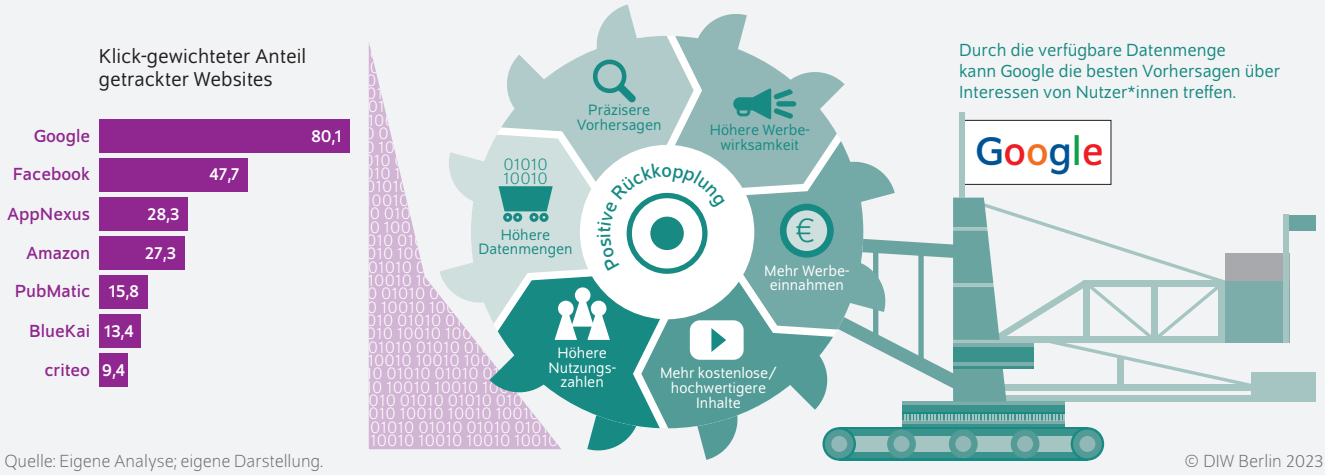
AUF EINEN BLICK

Daten als Wettbewerbsvorteil: Regulierer und Wettbewerbsbehörden sollten aufmerksam bleiben

Von Hannes Ullrich, Jonas Hannane, Tomaso Duso, Christian Peukert und Luis Aguiar

- Internetnutzer*innen können automatisiert beobachtet werden, um Profile für zielgerichtete Werbung zu erstellen
- Digitalunternehmen profitieren von Datenmengen – auch wenn der zusätzliche Nutzen weiterer Daten abnimmt
- Möglicher Netzwerkeffekt: Mehr Daten führen zu präziserer Personalisierung von Inhalten und Werbung, diese führen zu mehr Nutzer*innen, von denen sich zusätzliche Daten ergeben
- Google-Konzern Alphabet kann durch die Mengen an Daten, die er sammelt, einen Wettbewerbsvorteil erlangen
- Regulierung kann nötig sein, um zukünftige Markteintrittsschranken zu verhindern

Onlinekonzerne profitieren von Daten, aber Google sticht heraus: Daten könnten dem Google-Betreiber Alphabet eine uneinholbare Position im Markt bescheren



Quelle: Eigene Analyse; eigene Darstellung.

ZITAT

„Seinen Vorsprung konnte Google durch seine Pionierposition bekommen, und auch, weil Behörden spät erkannt haben, dass Daten ein zentraler Wettbewerbsfaktor sind. Mit Blick auf neue datenbasierte Technologien wie ChatGPT liegt nahe: Frühzeitige Regulierung kann helfen, den Markt offen für neue Angebote zu halten.“
— Hannes Ullrich —

MEDIATHEK

Audio-Interview mit Hannes Ullrich
www.diw.de/mediathek

Daten als Wettbewerbsvorteil: Regulierer und Wettbewerbsbehörden sollten aufmerksam bleiben

Von Hannes Ullrich, Jonas Hannane, Tomaso Duso, Christian Peukert und Luis Aguiar

ABSTRACT

Das umfangreiche Sammeln persönlicher Daten im Internet hat in den letzten zwei Jahrzehnten einen großen Markt für personalisierte Inhalte ermöglicht. Tausende kleine und große Unternehmen arbeiten an der Sammlung und Zusammenführung von Daten über das Nutzer*innenverhalten auf Websites, um über Eigenschaften und Interessen von Nutzer*innen zu lernen. Ganz an der Spitze befinden sich digitale Unternehmen wie Alphabet (Google), Amazon und Meta (Facebook), die den Großteil ihrer Umsätze durch Online-Werbung erwirtschaften. Diese Studie zeigt empirisch, dass eine steigende Menge an Daten zu einer besseren Vorhersagequalität von Personenmerkmalen führt, die für personalisierte Inhalte essenziell sind. Dennoch nehmen die Verbesserungen der Vorhersagequalität ab. Je mehr Daten dazukommen, desto weniger zusätzlichen Nutzen haben sie. Man kann hier von abnehmenden Skalenerträgen sprechen. Google, das bisher Klick-gewichtet auf über 80 Prozent der gängigsten Websites Nutzer*innenverhalten beobachten konnte, erreichte, bei gleichzeitig steigenden Nutzer*innenzahlen und beobachteten Websitebesuchen, jedoch signifikant weniger stark abnehmende Skalenerträge als seine Wettbewerber. Wenn aber Google allein von weiteren Daten durchweg mehr profitieren kann als seine Wettbewerber, können selbst innovative kleinere Unternehmen mangels Daten nur geringe Chancen haben, konkurrenzfähige Produkte zu vermarkten. Es würde sich eine uneinholbare Marktposition für den Google-Konzern ergeben, was zu Nachteilen für Konsument*innen führt.

Daten sind ein zentraler Wettbewerbsfaktor in digitalen Märkten. Sie ermöglichen innovative Dienstleistungen und Produkte von großem Wert für Konsument*innen. Neben Erfolgsmodellen wie der Google-Suchmaschine oder Amazons Online-Shop haben zuletzt große datenbasierte Sprachmodelle (Large Language Models, LLM) viel Aufmerksamkeit erregt. Sie werden zum Beispiel in ChatGPT vom Unternehmen OpenAI und auch in Produkten von Microsoft angewendet. Die große Bedeutung von Datenmengen wird jedoch zu einem gesellschaftlichen und wettbewerbsrechtlichen Problem, wenn der Zugang zu Daten einzelnen Unternehmen einen so großen Qualitäts- oder Kostenvorsprung verschafft, dass es zu Monopolisierung und Marktzutrittsschranken kommt, wodurch Märkte für neue Wettbewerber verschlossen werden.

Das kann bedeuten, dass nicht nur andere Unternehmen, sondern auch Konsument*innen Nachteile haben. Wenn Märkte hohe Zutrittsschranken haben, bleiben sie potenziell für Innovationen unzugänglich. Für Endnutzer*innen kann dies bedeuten, dass ihnen mittelfristig etwa neue technische Möglichkeiten, Verbesserungen in der Nutzungsfreundlichkeit oder gänzlich neue Online-Dienste, die sich zu besseren Alternativen bestehender Dienste entwickeln können, vorenthalten bleiben.

In der Tat deuten mehrere theoretische Studien darauf hin, dass Unternehmen in digitalen Märkten durch ihre Datenmengen Kipppunkte erreichen können, aus denen unüberwindbare Barrieren zum Markteintritt für andere Unternehmen folgen – selbst, wenn es anfangs nur kleine Unterschiede zwischen den Unternehmen gab.[1] Deshalb drehen sich viele politische Debatten darum, wie der Zugang zu Daten in Plattformmärkten reguliert werden kann, zum Beispiel im Digital Markets Act (DMA) der Europäischen Union.[2]

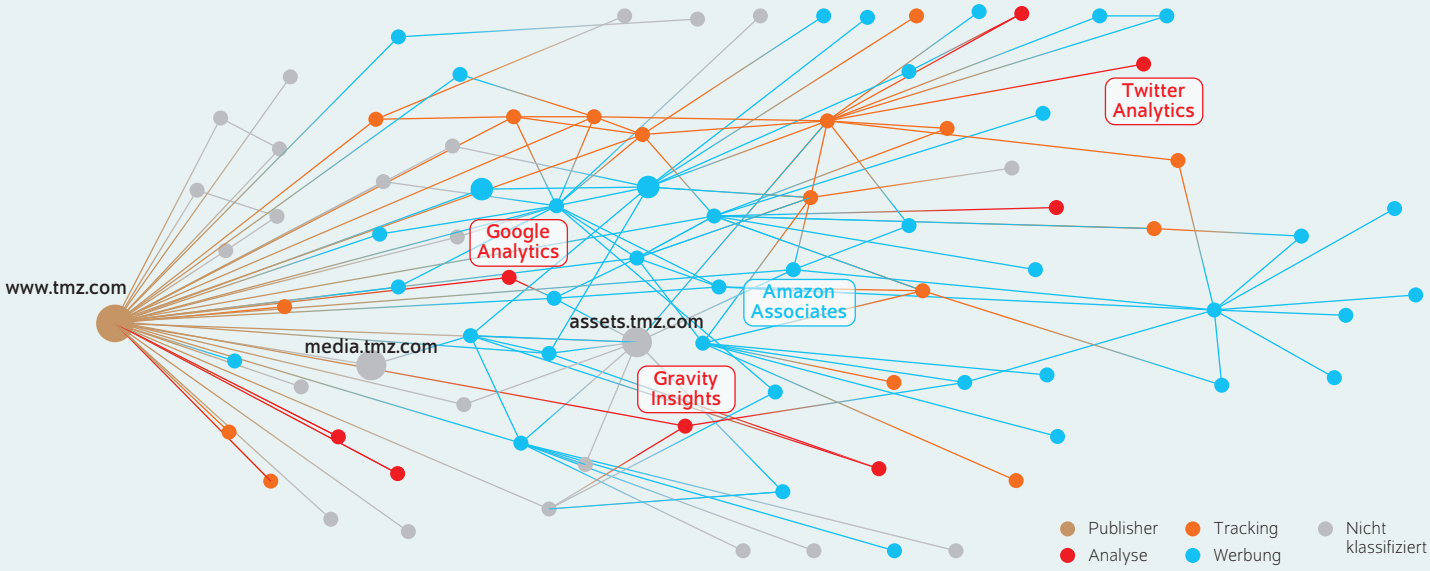
1 Vgl. Jan Krämer und Daniel Schnurr (2022): Big Data and Digital Markets Contestability: Theory of Harm and Data Access Remedies. Journal of Competition Law & Economics 18(2), 255–322. Jens Prüfer und Christoph Schottmüller (2021): Competing with Big Data. Journal of Industrial Economics 69, 967–1008. Bundeskartellamt (2017): Big Data und Wettbewerb (online verfügbar, abgerufen am 16. Juni 2023. Dies gilt auch für alle anderen Online-Quellen dieses Berichts, sofern nicht anders vermerkt).

2 Vgl. Digital Markets Act der Europäischen Union (online verfügbar).

DOI: https://doi.org/10.18723/diw_wb:2023-27-1

Abbildung 1

Tracker-Netzwerk am Beispiel der Website eines US-Boulevardmediums



Quelle: Eigene Analyse; eigene Darstellung.

Verschiedene Unternehmen sammeln Daten auf Websites durch Produkte mit unterschiedlichen Zwecken. Diese Daten werden an andere Unternehmen weitergegeben oder verkauft.

In der Studie, auf der dieser Wochenbericht basiert, wird überprüft, ob das großflächige Sammeln von Daten über das allgemeine Surfverhalten von Internetnutzer*innen den größten Online-Werbeplattformen einen unüberwindbaren Wettbewerbsvorsprung verleihen kann.[3]

Daten sind das wirtschaftliche Fundament von Tech-Giganten

Bekannte Unternehmen wie Alphabet (Anbieter von Google Search, Google Maps, Google Mail, Google Analytics, Google Play, Android und zahlreicher weiterer Produkte), Meta (Facebook, Instagram, WhatsApp), Microsoft (Windows, Office, Bing, Azure) und Amazon (Einzelhandel, Amazon Web Services, Amazon Ads) betreiben eigene Ökosysteme von vielen unterschiedlichen Produkten und Plattformen, in welchen Nutzungsdaten generiert werden.[4] Ihr Geschäftsmodell fußt unter anderem darauf, diese Daten auf unterschiedliche Arten und Weisen zu verknüpfen, zu nutzen und zu monetarisieren.

Daten sind wertvoll für Unternehmen, denn sie ermöglichen eine Verbesserung von digitalen Inhalten, zum Beispiel anhand von Empfehlungs-Algorithmen für Video- und Audiomedien sowie vielen anderen Online-Diensten, aber auch von Softwareprodukten und mehr. So können digitale Plattformen vielseitige datenbasierte Produkte anbieten, von denen Konsument*innen zunächst profitieren.

Online-Werbung ist das Paradebeispiel eines Produkts, bei dem Daten den wichtigsten Input darstellen. Für einige der größten digitalen Unternehmen wie Alphabet, Meta und zunehmend auch Amazon ist dieses Produkt von zentraler Bedeutung. Ihre Gewinne stammen zu einem großen Teil aus dem Verkauf von Werbung auf ihren Online-Plattformen wie Instagram, Facebook und Google Search oder auf Websites, die sie nicht selbst betreiben.[5] Die Werbeeinnahmen von Alphabet und Meta zusammen beliefen sich im Jahr 2022 auf 338 Milliarden US-Dollar.[6] Dieser Erfolg wird unter anderem der Fähigkeit dieser Unternehmen zugeschrieben, so zielgenau Werbung zu schalten, dass sie bei Konsument*innen häufiger zum Klick oder direkt zum Kauf führt.

Datenmengen können bestimmend sein

Gezielte Werbung basiert auf umfangreichen Daten, die von Unternehmen gesammelt und analysiert werden. Diese Daten entstehen häufig als Nebenprodukt bei der Nutzung von digitalen Anwendungen oder beim Surfen im Internet. Hierbei können zum Beispiel Daten über das verwendete

3 Vgl. Luis Aguiar et al. (2023): Returns to Web Tracking Data, mimeo (online verfügbar).

4 In der restlichen Publikation wird der Markenname Google statt der Name des Trägerkonzerns Alphabet benutzt, da es zu großen Teilen um Google-Dienste geht und dies die deutlich bekanntere Bezeichnung ist. Weiterhin wird im Zusammenhang mit der Untersuchung Facebook genannt, da die Stichprobe aus der Zeit stammt, in der das Unternehmen noch nicht Meta hieß.

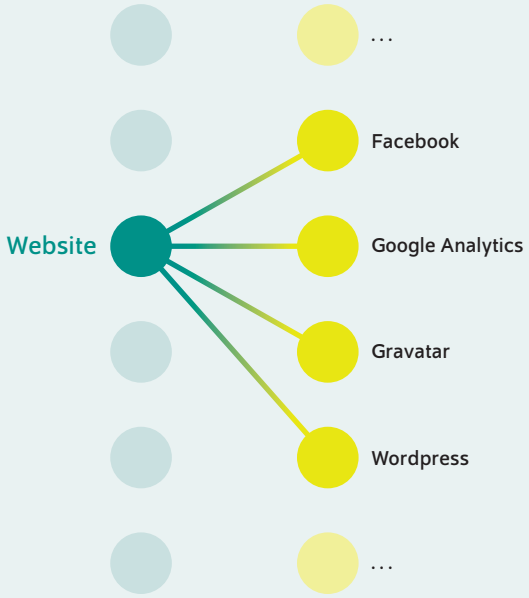
5 Vgl. Competition and Markets Authority (2020): Online platforms and digital advertising market study (online verfügbar).

6 Vgl. Darstellung von Statista (online verfügbar).

Abbildung 2

Verschiedene Tracker im HTML-Code einer Website

```
…
<div id="rr social widgets facebook">
        <iframe src="// www.facebook.com /plugins/
                like.php?href=http%3A%2Fwww.facebook.com
                %2Ftechcrunch&amp;layout=standard…"…"
        </iframe>
</div>
…
<script type="text/javascript">
        var gaJsHost = (("https:" == document.
                location.protocol)?"https://ssl.":"http://www.");
        document.write(unescape("%3Cscript
                src='" + galsHost +
                " google-analytics.com /ga.js'
                type='text/javascript'%3E%3C/script%3E"));
</script>
…
<script src='http:// s.gravatar.com /js/
        type='text/javascript'
        gprofiles.js?aa&#038;ver=3.5-alpha-21304'>
</script>
…
<noscript>
        <img src="http:// stats.wordpress.com /
                b.gif?v=noscript" … />
</noscript>
…
```



Quelle: Eigene Analyse; eigene Darstellung.

Websites laden zahlreiche Tracker verschiedener Anbieter, die von den Seitenbetreibern in den Quellcode eingebaut werden.

Gerät, insbesondere aber auch die Adressen einzelner besuchter Websites erhoben werden. Neben bekannten Unternehmen wie Alphabet und Meta sammeln mehr als 4000 eher unbekannte Datenbroker ebendiese Daten (Abbildung 1). Diese Unternehmen konkurrieren im Internet, um mit Analyse-, Technologie- oder anderen Dienstleistungen auf möglichst vielen und relevanten Websites integriert zu werden. Sie bieten also Leistungen, die für Websitebetreiber Vorteile bieten. Teil dieses Handels ist dann, dass die Unternehmen dort die Klicks von Nutzer*innen beobachten und, auf dem Klickverhalten basierend, personenspezifische Profile erstellen können.[7] Demografische Profile, zum Beispiel eine Unterteilung von Personen nach Alter, Geschlecht oder Wohnort, werden dann verwendet, um individuelle Interessen als Grundlage für personalisierte Werbung abzuleiten.[8] Werbeanzeigen können so für gewünschte Zielgruppen verkauft werden.

Das Ergebnis aus Sicht der Nutzer*innen ist aus der Internetsuche bekannt. Wenn man bei Google Search oder Bing nach „Herrentasche“ sucht, ist die Wahrscheinlichkeit hoch, dass in der Ergebnisliste eine Werbeanzeige für eine Handtaschenmarke erscheint. Welche Herrentaschen man zu sehen bekommt, kann wiederum von der Art des eigenen Profils abhängen. Diese Profile werden erstellt, indem Google, Meta und andere Datenbroker auf vielen Websites kleine Codeschnipsel platzieren (Abbildung 2), die Informationen über die Nutzer*innen – zum Beispiel von welcher Website sie kommen, wie lange sie auf der Seite bleiben, welche Links sie anklicken – an die Server der trackenden Unternehmen senden können. Durch das Speichern von sogenannten Cookies auf den lokalen Rechnern von Nutzer*innen können diese über Websites hinweg identifiziert werden.[9] Anhand dieser Informationen können Unternehmen die Charakteristika der Nutzer*innen abschätzen. Zum Beispiel wird ein Nutzer, der häufig die Website eines Sportmagazins besucht, das auf eine junge Kundschaft spezialisiert ist, bei der Suche nach dem Begriff „Herrentasche“ eher ein Produkt für 20-Jährige als für 60-Jährige finden.

Datennetzwerkeffekte machen Profile präziser

Um auf dem Markt für personalisierte Werbung zu bestehen, können zum einen ausreichende Datenmengen und zum anderen technologische Innovationen in der Verwertung der Daten notwendig sein.

7 Vgl. Hannes Ullrich et al. (2022): Plattformen wie Facebook können mehr als die Hälfte der Internetaktivität beobachten. DIW Wochenbericht Nr. 29/30, 400–406 (online verfügbar).

8 Vgl. Nico Neumann, Catherine E. Tucker und Timothy Whitfield (2019): Frontiers: How effective is third-party consumer profiling? Evidence from field studies. Marketing Science 38 (6), 918–926. Jon Keegan und Joel Eastwood (2023): From “Heavy Purchasers” of Pregnancy Tests to the Depression-Prone: We Found 650,000 Ways Advertisers Label You. The Markup, 8. Juni 2023 (online verfügbar).

9 Bei First-Party-Cookies geschieht dies durch den Anbieter der besuchten Website. Werden diese von Drittanbietern geladen und Daten außerhalb der besuchten Website verknüpft, so spricht man von Third-Party-Cookies. Die verpflichtende Einholung der Zustimmung zur Nutzung von Third-Party-Cookies ist ein Hauptpunkt der seit Mai 2018 geltenden EU-Datenschutz-Grundverordnung (EU-DSGVO).

Die Profile werden von Datenbrokern und Plattformen mit Methoden des maschinellen Lernens erstellt. So werden aus den von ihnen gesammelten Daten Vorhersagen über die Merkmale der Profilinhaber*innen getroffen. Eine Besonderheit in der Auswertung von Online-Nutzungsdaten sind sogenannte Datennetzwerkeffekte, welche eine sehr hohe Attraktivität und Qualität von digitalen Diensten und Produkten ermöglichen können.

Datennetzwerkeffekte entstehen durch eine positive Rückkopplung: Präzisere Vorhersagen ermöglichen ein besseres Erreichen der Zielgruppe durch Werbetreibende. Dies erhöht die Wirksamkeit der Werbung, wodurch im Gegenzug Online-Plattformen, auf denen die Werbung geschaltet wird, höhere Preise für die Bereitstellung der Werbefläche erzielen. Die höheren Einnahmen der Plattformen ermöglichen diesen, mehr kostenlose oder hochwertigere Inhalte bereitzustellen. Dies lockt wiederum mehr Nutzer*innen auf die Plattform, die so mehr Daten erhält. Die höheren Datenmengen können wiederum die Präzision der Vorhersagen verbessern, wodurch sich der Zyklus fortsetzt (Abbildung 3).

In zweiseitigen Märkten, in denen sich zwei Kund*innengruppen in ihrem Nachfrageverhalten gegenseitig beeinflussen, etwa Internetnutzer*innen und Werbetreibende, ist häufig eine Winner-takes-all-Dynamik zu beobachten, welche zum Kippen dieser Märkte führt. Genauso können auch Datennetzwerkeffekte zu Monopolisierungstendenzen beitragen.

Zusätzliche Daten ermöglichen zielgenauere Werbung, der zusätzliche Ertrag nimmt aber ab

Die Relevanz solcher Datennetzwerkeffekte hängt vom Wert wachsender Datenmengen für die Produktqualität und damit für den Unternehmenserfolg im Markt ab. Es liegt auf der Hand, dass die Genauigkeit der Vorhersagen zunimmt, je mehr Daten einem trackenden Unternehmen zur Verfügung stehen: Daten weisen sogenannte Skalenerträge auf (Kasten 1). Für einzelne Anwendungen von Internet- und Produktsuchmaschinen wurde in bestehenden Studien gezeigt, dass Skalenerträge von Daten abnehmend sind.[10] Wenn ein trackendes Unternehmen Informationen über weniger Nutzer*innen hat, ist der Wert, eine*n einzelne*n zusätzliche*n Nutzer*in zu beobachten, groß. Der Wert zusätzlicher Daten ist aber viel kleiner, wenn schon Daten über Millionen von Nutzer*innen vorhanden sind. Das hat grundlegende Auswirkungen auf den Wettbewerb und die Entwicklung eines Marktes.

Abbildung 3

Datennetzwerkeffekte

Quelle: Eigene Darstellung.

Mehr Daten können zu mehr Einnahmen, mehr Nutzer*innen und daraus folgend zu noch mehr Daten führen.

Lässt sich schon mit sehr geringen Datenmengen eine hohe Vorhersagekraft erreichen, so ist es auch für kleine Konkurrenten mit wenigen Daten leicht, ein erfolgreiches Produkt anzubieten. Unterscheidet sich aber der Wert zusätzlicher Daten systematisch zwischen konkurrierenden Firmen, so können Datennetzwerkeffekte zu problematischen Marktstrukturen führen.

Durch den Vergleich von Skalenerträgen über viele Unternehmen hinweg kann in der Studie analysiert werden, ob die Verfügbarkeit von Daten zu einer möglichen Markteintrittsbarriere für konkurrierende Unternehmen werden kann. Das wäre dann der Fall, wenn die Skalenerträge eines Unternehmens, das Tracker auf den meisten Websites installiert hat und daher viele und vielfältige Daten sammelt, in Bezug auf die Anzahl der Nutzer*innen und Websites weniger stark abnehmen als für andere, kleinere Konkurrenten.

Um diese Fragestellung zu beantworten, kommen Methoden des maschinellen Lernens zum Einsatz, die es ermöglichen, das Nutzungsverhalten im Internet, beispielsweise regelmäßige Besuche auf bestimmten Websites, mit persönlichen Eigenschaften wie dem Alter oder Geschlecht der Nutzer*innen statistisch zu verknüpfen.

10 Vgl. Di He et al. (2017): Scale effects in web search. In Web and Internet Economics, WINE 2017 Proceedings, 294–310. Patrick Bajari et al. (2019): The impact of big data on firm performance: An empirical investigation. AEA Papers & Proceedings 109, 33–37. Tobias Klein et al. (2023): How Important Are User-generated Data For Search Result Quality? Experimental Evidence. CEPR Discussion Paper DP17934. Christian Peukert, Ananya Sen und Jörg Claussen (2023): The Editor and the Algorithm: Recommendation Technology in Online News. Im Erscheinen bei Management Science. Eine Studie findet lokal zunehmende Skalenerträge von Daten für die Qualität einer Suchmaschine und diskutiert die damit einhergehenden Konsequenzen für den Wettbewerb. Vgl. Maximilian Schaefer und Geza Sapi (2023): Complementarities in Learning from Data: Insights from General Search, mimeo (online verfügbar).

Kasten 1

Skalenerträge für datenbasiertes Lernen

Die zentrale Frage der Studie lautet, wie sehr sich die Genauigkeit der von Datenbrokern und anderen Unternehmen, auf Vorhersagen basierten, erstellten Konsument*innenprofile verändert, je nachdem wieviel Daten zur Verfügung stehen. In der Volkswirtschaftslehre spricht man von sogenannten Skalenerträgen, um die Abhängigkeit einer Produktionsmenge, in dieser Studie also die Vorhersagequalität, von der Menge eingesetzter Produktionsfaktoren, das sind hier die Daten, zu beschreiben. Dabei unterscheidet man zwischen konstanten, zunehmenden oder abnehmenden Skalenerträgen, je nachdem ob die Produktionsmenge bei einer proportionalen Erhöhung der Produktionsfaktoren um einen Faktor μ ebenfalls um μ ansteigt, mehr als μ ansteigt oder weniger als μ ansteigt (Abbildung).

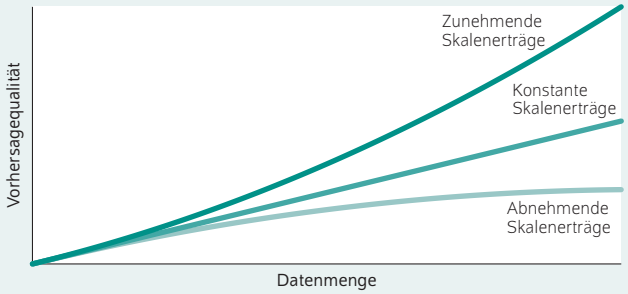
Eine Quantifizierung von datenbasierten Skalenerträgen ist essenziell, um ein besseres Verständnis über die Dynamiken in zweiseitigen Märkten zu erhalten. Aufgrund der durch Netzwerkeffekte generierten positiven Rückkopplung können nicht- oder nur schwach abnehmende Skalenerträge zu Markteintrittsbarrieren und Monopolisierung von Märkten führen. Dadurch wurde ein Eingreifen von Wettbewerbsbehörden oft motiviert.

Daten über Internetverläufe können entlang von zwei Dimensionen gesammelt werden: Es können sowohl mehr Daten über mehr Internetnutzer*innen gesammelt werden als auch mehr Daten über die einzelnen Internetnutzer*innen. Die Studie analysiert die Skalenerträge aus Daten und zeigt, wie sich die Vorhersagequalität verbessert, wenn mehr Daten in jeder Dimension für sich genommen verfügbar sind und ob es eine Wechselwirkung zwischen beiden Dimensionen gibt.

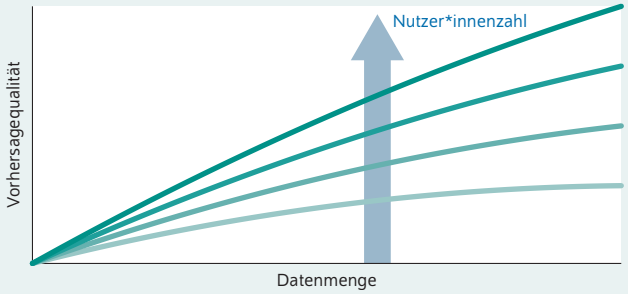
Abbildung

Skalenerträge

Mögliche Formen von Skalenerträgen



Beispielhaft: Unterschiedlich abnehmende Skalenerträge in Abhängigkeit der Anzahl von Nutzer*innen



Quelle: Eigene Darstellung.

Grundlage der Analyse sind Daten einer repräsentativen Stichprobe der Marktforschungsfirma Nielsen aus dem Jahr 2016. In einer Erhebung haben 4989 Internetnutzer*innen in den USA persönliche Merkmale angegeben und ihr Surfverhalten über ein Jahr speichern lassen, indem ein Browser-Plugin den gesamten Verlauf aller von den individuellen Nutzer*innen besuchten Websites aufzeichnete. Hierbei wurden sowohl die Anzahl an getätigten Klicks pro Website als auch die Verweildauer auf der entsprechenden Seite aufgenommen. Aus einem zweiten Datensatz der Initiative HTTP Archive wird für jede besuchte Website die Information übernommen, welche externen Dienstleister beim Laden der Website Informationen übertragen konnten.[11] Durch das Verknüpfen dieser beiden Datenquellen kann der durch individuelle externe Dienstleister beobachtbare Anteil des Surfverhaltens für alle Internetnutzer*innen quantifiziert werden.

Basierend auf den jeweiligen Daten zu besuchten Websites, über die jedes einzelne Unternehmen verfügt, wird in der Studie ein Algorithmus zur Klassifizierung von Personen nach zehn Eigenschaften angewandt: Alter in vier Kategorien, Geschlecht, hohes Einkommen, hohes Bildungsniveau, Beschäftigungsstatus, Kinder im Haushalt und Lokalisierung in einem wahlentscheidenden US-Bundesstaat. Um den Wert zusätzlicher Daten für die Qualität der Klassifikation zu messen, wird stets derselbe Algorithmus verwendet, während die Menge der Daten variiert (Kasten 2).[12] Das sind die Anzahl der beobachtbaren und besuchten Websites, gewichtet durch die Anzahl an Klicks auf einer Website, sowie die Anzahl an beobachteten Nutzer*innen.

Aus der Analyse wird deutlich, dass die Qualität der Klassifizierung mit zunehmenden Daten steigt. Der Zuwachs der Qualität wird aber immer kleiner. Am Beispiel der Variablen Geschlecht zeigt sich: Sogar mit wenigen Daten, etwa mit nur 20 Prozent der klick-gewichteten Websites im betrachteten Analysedatensatz, ist eine gute Klassifizierung des Geschlechts möglich. Es würden ungefähr

11 Vgl. HTTP Archive (online verfügbar).

12 Der verwendete Algorithmus XGBoost wird auch in der Unternehmenspraxis für vergleichbare Problemstellungen angewandt. Vgl. zum Beispiel die Website Kaggle (online verfügbar).

Kasten 2

Maschinelles Lernen und das Messen unternehmensspezifischer Skalenerträge

Maschinelles Lernen erlaubt Vorhersagen basierend auf Korrelationen eines vorherzusagenden Ergebnisses. Das kann beispielsweise die Zugehörigkeit einer Person zu einer bestimmten demografischen Gruppe sein. Für die Vorhersagen werden beobachtbare Informationen genutzt – im Falle dieser Studie sind das Internetnutzungsverläufe. Das Erstellen von Profilen für das Bereitstellen personalisierter Werbung entspricht sogenannten Klassifizierungsverfahren, in denen versucht wird, Internetnutzer*innen mittels maschinellen Lernens in verschiedene werberelevante Kategorien (zum Beispiel „weiblich" als Konsumentinnenprofil) einzuteilen. In der Studie wird *extreme gradient boosting*, XGBoost, angewandt, ein gängiger Algorithmus für strukturierte Datensätze, welcher auf einer Ansammlung verschiedener Klassifizierungsmodelle basiert.

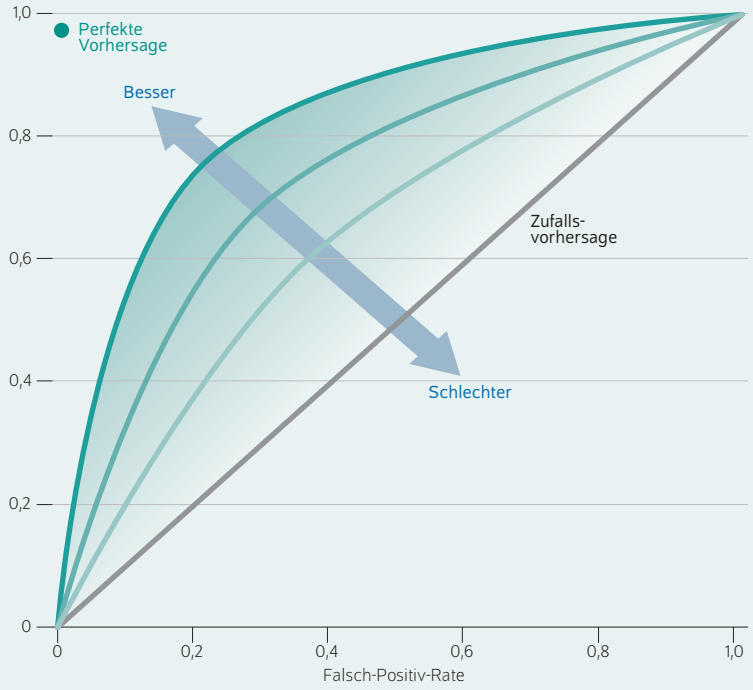
Die *Area Under the ROC Curve* (AUC) ermöglicht es, die Vorhersagequalität verschiedener Klassifikatoren anhand einer Kennzahl zu bemessen. Auf der *Receiver Operating Characteristic* (ROC)-Kurve wird die Richtig-Positiv-Rate (der Anteil richtig positiv klassifizierter Fälle aus allen positiven Fällen), der Falsch-Positiv-Rate (der Anteil falsch positiv klassifizierter Fälle aus allen nichtpositiven Fällen) einer Klassifizierungstechnologie gegenübergestellt. Die ROC-Kurve stellt alle machbaren Abwägungen zwischen der Richtig-Positiv-Rate und der Falsch-Positiv-Rate der Technologie dar. Die AUC liegt zwischen 0,5 (reine Zufallsklassifikation) und 1 (bestmögliches Klassifizierungsverfahren).[1]

Um firmenspezifische Skalenerträge sowohl über die Anzahl von Nutzer*innen als auch über die Anzahl an beobachtbaren Websites zu messen, erstellt die Studie hypothetische Teildatensätze. Für jedes der zehn Klassifikationsprobleme wird hierfür die Detailtiefe (die Anzahl der Spalten) sowie die Größe (die Anzahl der Zeilen) des einem jeweiligen Datenbroker zu Verfügung stehenden Datensatzes systematisch variiert. Letzterer entspricht stets einem tabellarischen Datensatz, in welchem jede Zeile einem*einer Internetnutzer*in und jede Spalte einer Website zugeordnet ist. Zur Erstellung der Teildatensätze werden zufällig ausgewählte Internetnutzer*innen und Websites aus dem Datensatz eines Datenbrokers entfernt. Dies geschieht basierend auf einem vorher definierten Raster. Beispielsweise werden für große Datenbroker, welche über 15 000 Webseiten tracken, pro Klassifikationsaufgabe jeweils ein Zehntel vom Ganzen der Internetnutzer*innen und Websites entfernt.

Auf Basis dieser Teildatensätze wird eine Simulationsanalyse durchgeführt, bei der für jede Klassifikationsaufgabe der Algorithmus zur Vorhersage einer Eigenschaft der Internetnutzer*innen angewandt wird. Die dabei erzielten Vorhersagequalitäten, gemessen durch die AUC, werden gesammelt und gespeichert. Dies ermöglicht die anschließende Quantifizierung von firmenspezifischen Skalenerträgen anhand von Regressionsanalysen.

Abbildung

ROC-Kurve als Maß für Vorhersagequalität
Korrekt-Positiv-Rate



Anmerkungen: Beispielhafte Illustration. Die Korrekt-Positiv-Rate (Falsch-Positiv-Rate) bezeichnet den Anteil der korrekt (fälschlicherweise) positiv klassifizierten Fälle. Je weiter links oben die ROC-Kurve verläuft, desto besser ist die Vorhersagequalität. Die Diagonale ist die ROC-Kurve einer reinen Zufallsklassifikation (Münzwurf).

Quelle: Eigene Darstellung.

1 Shan Huang, Michael Allan Ribers, Hannes Ullrich (2021): Der gesellschaftliche Mehrwert verknüpfter Daten: Algorithmen als Entscheidungshilfen bei Antibiotikaverschreibungen. DIW Wochenbericht Nr. 13/14, 239–246 (online verfügbar).

95 Prozent der bestmöglichen Klassifizierung anhand von allen Daten erreicht werden. Das bedeutet: Auch mit eher wenigen Daten können die verschiedenen Dienste relativ gut Eigenschaften der Nutzer*innen anhand deren Verhaltens vorhersagen. Um eine weitere deutliche Steigerung der Klassifizierungsqualität zu erreichen ist aber ein so hoher Anteil an getrackten Websites nötig, wie ihn nur Google erreicht (Abbildung 4).

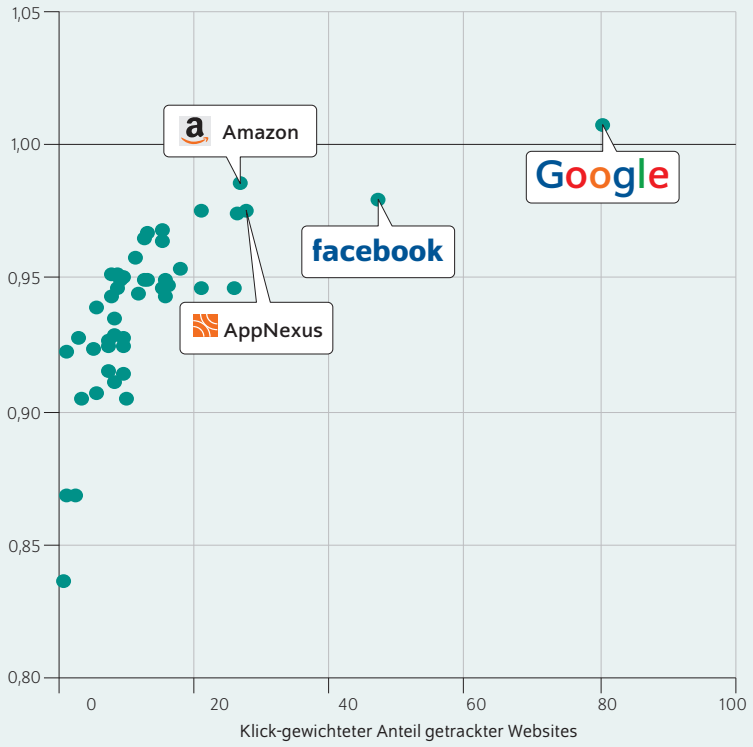
Google hat einen Wettbewerbsvorteil

Um Skalenerträge von Daten für jedes Unternehmen zu schätzen, wurde eine Simulationsanalyse durchgeführt, in der die Zahl der beobachteten Nutzer*innen und der beobachteten Websites variiert und mit der damit erreichbaren Klassifizierungsqualität verknüpft wird (Kasten 2). Hierauf basierend wird in einer Regressionsanalyse der

Abbildung 4

Vorhersagequalität der demografischen Profilbildung verschiedener Plattformen

Vorhersagequalität; Klick-gewichteter Anteil getrackter Websites in Prozent



Quelle: Eigene Berechnungen.

Die Datenmengen von Google führen zu einem Vorsprung bei der Vorhersagequalität der Nutzer*innenprofile.

Zusammenhang zwischen der Klassifizierungsqualität und der Anzahl der beobachteten Nutzer*innen sowie der beobachteten Websites geschätzt.

Sowohl für alle Firmen zusammengenommen als auch für die größten Firmen Google, Facebook, AppNexus und Amazon separat werden abnehmende Skalenerträge in der Anzahl an Nutzer*innen und Websites gemessen (Tabelle 1).

Die Ergebnisse für Google stechen jedoch heraus: Der Interaktionseffekt zwischen der Anzahl der Nutzer*innen und Websites hat einen positiven, statistisch signifikanten Koeffizienten. Das deutet darauf hin, dass Skalenerträge in der Anzahl an Nutzer*innen und Websites für Google weniger stark abnehmen als für andere Firmen. Somit können Daten allein einen nur schwer einholbaren Wettbewerbsvorteil für Google darstellen.

Fazit: Daten begünstigen eine konzentrierte Marktstruktur – Regulierung kann nötig sein

Marktführende Digitalunternehmen verfügen über enorme Datenmengen zu Verhalten und Vorlieben der Verbraucher*innen, die sie zur Verbesserung ihrer Produkte und Dienstleistungen nutzen können. Durch den Zugang zu umfangreichen Datenbeständen können sie ihre Angebote zunehmend personalisieren, gezielte Werbung schalten und ihre Dienste kontinuierlich verbessern. Dies schafft einen Wettbewerbsvorteil gegenüber kleineren Unternehmen, die nicht über vergleichbare Ressourcen verfügen, um solche Daten zu sammeln und zu nutzen.

Die Untersuchung zeigt einen möglichen systematischen Unterschied in der Qualität der Bildung von Werbeprofilen von Unternehmen, die Nutzungsdaten sammeln. Die Ergebnisse suggerieren somit, dass das Teilen von Daten zwischen Online-Diensten ein wirksames Instrument sein kann, um Wettbewerb zu fördern und Innovationsanreize zu stärken. Der Digital Markets Act (DMA) der EU sieht vor, dass sogenannte Gatekeeper wie Alphabet und Meta, definiert auf Basis der Umsatzhöhe und Anzahl an Nutzer*innen, verpflichtet werden, Daten, die während der Nutzung ihrer Kerndienste entstehen, an Endnutzer*innen sowie gewerbliche Nutzer weiterzugeben.[13] Der Durchsetzung dieser Regelung kommt daher für die Wettbewerbsdynamik insbesondere in der EU eine große Bedeutung zu.

Grundsätzlich wird die Profilbildung anhand von Third-Party-Cookies bereits durch die Datenschutzgrundverordnung der EU (EU-DSGVO) und zukünftig auch durch den DMA insbesondere in Europa zunehmend erschwert. Dennoch arbeiten die marktführenden Akteure der Online-Werbebranche an technologischen Entwicklungen, die Personalisierung und Kontextualisierung von Inhalten zwar datenschutzkonform ermöglichen, ihre Marktstellung bei den gegebenen Marktstrukturen aber weiter zementieren können. Dies betrifft nicht nur datenbezogene Strategien, sondern auch andere strategische Entscheidungen der marktbeherrschenden Plattformen, die zu einer Abschottung in Teilmärkten führen können.

Genau deswegen hat die Europäische Kommission zuletzt im Juni 2023 in einem bis dato einzigartigen Schritt angekündigt, auf die Marktstruktur abzuzielen und eine Zerschlagung der integrierten Werbedienste von Google zu erwirken, um den Wettbewerb von anderen Anbietern von Werbetechnologiediensten, Werbetreibenden und Online-Publishern zu ermöglichen.[14]

Diese Überlegungen sind nicht nur für den hier analysierten Markt von großer Relevanz, sondern im Allgemeinen für die aktuelle Wettbewerbsdynamik zwischen digitalen Plattformen in neu entstehenden Märkten. Zuletzt haben beispielsweise die dominierenden digitalen Unternehmen viele Ressourcen in die Entwicklung von großen Sprachmodellen

13 Gemäß dem DMA sind die Kerndienste von Online-Plattformen soziale Medien, Suchmaschinen, Videoplattformen, Kommunikationsdienste, intermediäre Dienstleistungen, Cloud-Dienste, Betriebssysteme und Werbe-Dienste. Auch der sich in der Abstimmung befindende europäische Data Act soll Unternehmen verpflichten, Daten mit anderen Unternehmen und Kund*innen zu teilen.

14 Pressemitteilung der Europäischen Kommission vom 14. Juni 2023: Antitrust: Commission sends Statement of Objections to Google over abusive practices in online advertising technology (online verfügbar).

Tabelle 1

Regressionsergebnisse der Simulationsanalyse

Variablen	(1) Alle Anbieter	(2) Google	(3) Facebook	(4) AppNexus	(5) Amazon
Anzahl der Nutzer*innen (in Tausend)	0,085***	0,084***	0,087***	0,041***	0,083***
Anzahl der Nutzer*innen (quadriert)	−0,012***	−0,012***	−0,011***	−0,004***	−0,011***
Anteil der getrackten Websites	0,420***	0,370***	0,306***	0,868***	0,430***
Anteil der getrackten Websites (quadriert)	−0,466***	−0,642***	−0,388**	−0,519	−1,426***
Anzahl der Nutzer*innen × Anzahl der Websites	0,008	0,038***	−0,010	−0,089**	0,036

Anmerkung: Die Sternchen an den Werten bezeichnen das statistische Signifikanzniveau. Je mehr Sternchen, desto geringer die Irrtumswahrscheinlichkeit: ***, ** und * geben die Signifikanz auf dem Ein-, Fünf- und Zehn-Prozent-Niveau an.

Lesehilfe: Die Vorhersagequalität ist relativ zur bestmöglichen Vorhersagequalität basierend auf dem vollständigen Datensatz definiert. Der Durchschnitt der so definierten Vorhersagequalität ist 0,86, was sich durch Teilen der durchschnittlichen absoluten Vorhersagequalität von 0,63 durch den Durchschnitt der bestmöglichen Vorhersagequalität von 0,73 ergibt. Die Zeilen 1 und 3 isoliert betrachtet ist für alle Anbieter (Spalte 1) ein Anstieg der Anzahl an Nutzer*innen um Tausend (Zeile 1) mit einem Anstieg der Vorhersagequalität um 8,5 Prozentpunkte assoziiert. Die Erhöhung des Anteils getrackter Websites (Zeile 3) um einen Prozentpunkt ist für Facebook (Spalte 3) mit einem Anstieg der Vorhersagequalität um 0,306 Prozentpunkte assoziiert. Aufgrund der quadratischen Terme mit negativen Werten in Zeilen 2 und 4 nehmen diese positiven Assoziationen allerdings zunehmend ab. Für Google (Spalte 2) wird diese Abnahme gedämpft, da ein gleichzeitiger Anstieg der Anzahl an Nutzer*innen um Tausend (Zeile 1) und des Anteils getrackter Websites (Zeile 3) positiv zur Vorhersagequalität beiträgt.

Quelle: Eigene Berechnungen.

investiert, die mittlerweile für die breite Masse zugänglich gemacht wurden. Für die Qualität dieser Modelle sind Daten und die Validierung durch Menschen essentiell. OpenAI hat bereits intern menschliche Bewertungen von generierten Inhalten verwendet, um das Modell GPT-4, das für den ChatGPT-Dienst verwendet wird, zu verbessern.[15]

Somit ist denkbar, dass das frühzeitige Bereitstellen von GPT-unterstützten Programmen durch Google, Meta und OpenAI/Microsoft ein frühzeitiges Verbessern der Algorithmen dank des Testens und Validierens durch Millionen von Nutzer*innen erlaubt. Wer diese Gelegenheit trotz potenzieller Risiken frühzeitig nutzt, kann so in Plattformmärkten Wettbewerbsvorteile generieren und langfristig Marktmacht zementieren. Ein verstärktes Augenmerk auf diese Dynamik könnte die aktuellen Diskussionen um die Risiken von KI in konstruktivere und wirksamere Bahnen lenken als (wenn auch wichtige) Diskussionen über die, von manchen prophezeite, Verselbstständigung der KI.

Google und andere BigTech-Unternehmen konnten ihre Marktstellung und den hier gezeigten Wettbewerbsvorteil unter anderem gewinnen, weil diese Märkte über die letzten Jahrzehnte kaum ex-ante reguliert wurden und die ex-post wettbewerbsrechtliche Aufsicht nicht stark durchgesetzt wurde. Um in Märkten, für die die angesprochenen Sprachmodelle relevant sind, eine ähnliche Monopolisierung zu vermeiden, sollte am Beispiel der Entwicklung von Google gelernt werden. Frühe Eingriffe in diese Märkte können nötig werden, um Wettbewerbern den Zutritt zum Markt nicht zu versperren und weiterhin Innovationen zu ermöglichen.

15 Ajay Agrawal, Joshua Gans und Avi Goldfarb (2023): How Large Language Models Reflect Human Judgment. Harvard Business Review vom 12. Juni (online verfügbar).

Hannes Ullrich ist wissenschaftlicher Mitarbeiter der Abteilung Unternehmen und Märkte im DIW Berlin | hullrich@diw.de

Jonas Hannane ist Doktorand in der Abteilung Unternehmen und Märkte im DIW Berlin | jhannane@diw.de

Tomaso Duso ist Leiter der Abteilung Unternehmen und Märkte im DIW Berlin | tduso@diw.de

Christian Peukert war Fellow in der Abteilung Unternehmen und Märkte im DIW Berlin

Luis Aguiar ist Professor für Ökonomie an der Universität Zürich

JEL: D18, L40, L50, L86, M38

Keywords: Competition, Data Network Effects, Web Tracking, Platforms.

IMPRESSUM

DIW Berlin — Deutsches Institut für Wirtschaftsforschung e. V.
Mohrenstraße 58, 10117 Berlin
www.diw.de
Telefon: +49 30 897 89–0 Fax: –200
90. Jahrgang 5. Juli 2023

Vertrieb
DIW Berlin Leserservice, Postfach 74, 77649 Offenburg
leserservice@diw.de
Telefon: +49 781 639 67 20

Gestaltung
Roman Wilhelm, Stefanie Reeg, Eva Kretschmer, DIW Berlin

Umschlagmotiv
© imageBROKER / Steffen Diemer

Satz
Satz-Rechen-Zentrum Hartmann + Heenemann GmbH & Co. KG, Berlin

Druck
USE gGmbH, Berlin

ISSN 0012-1304; ISSN 1860-8787 (online)

Abonnieren Sie auch unseren DIW- und/oder Wochenbericht-Newsletter unter **www.diw.de/newsletter**