cerre Centre on Regulation in Europe

REPORT

September 2020

Jan Krämer
Daniel Schnurr
Sally Broughton Micova

THE ROLE OF DATA FOR DIGITAL MARKETS CONTESTABILITY

CASE STUDIES AND DATA ACCESS REMEDIES

The project, within the framework of which this report has been prepared, has received the support and/or input of the following organisations: ARCEP, BIPT, Facebook, Mediaset, Microsoft, Snap Inc. and Telefónica.

As provided for in CERRE's by-laws and in the procedural rules from its "Transparency & Independence Policy", this report has been prepared in strict academic independence. At all times during the development process, the research's authors, the Joint Academic Directors and the Director General remain the sole decision-makers concerning all content in the report.

The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE, or to any sponsor or members of CERRE.

info@cerre.eu
www.cerre.eu

Table of contents

About CERRE

Providing top-quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) promotes robust and consistent regulation in Europe's network and digital industries. CERRE's members are regulatory authorities and operators in those industries as well as universities.

CERRE's added value is based on:

- its original, multidisciplinary and cross-sector approach;
- the widely acknowledged academic credentials and policy experience of its team and associated staff members;
- its scientific independence and impartiality;
- the direct relevance and timeliness of its contributions to the policy and regulatory development process applicable to network industries and the markets for their services.

CERRE's activities include contributions to the development of norms, standards and policy recommendations related to the regulation of service providers, to the specification of market rules and to improvements in the management of infrastructure in a changing political, economic, technological and social environment. CERRE's work also aims at clarifying the respective roles of market operators, governments and regulatory authorities, as well as at strengthening the expertise of the latter, since in many Member States, regulators are part of a relatively recent profession.

About the authors

Jan Krämer is a CERRE Joint Academic Director and Professor for Information Systems at the University of Passau, Germany, where he holds the Chair for Internet & Telecommunications Business and is the director of the Passau International Centre for Advanced Interdisciplinary Studies. He has a diploma degree in Business Engineering and Management, and a Ph.D. in Economics, both from the Karlsruhe Institute of Technology. He has published numerous articles in the leading journals in the areas of Information Systems, Economics and Marketing. His current research interests include the regulation of telecommunications and Internet markets, as well as digital ecosystems and data-driven business models.

Daniel Schnurr leads the research group Data Policies at the University of Passau. He received his Ph.D. in Information Systems from the Karlsruhe Institute of Technology, where he previously studied Information Engineering and Management (B.Sc. & M.Sc.). He has published in leading journals in the areas of Information Systems and Economics on competition and cooperation in telecommunications markets, open access regulation as well as data sharing in digital markets. His current research focuses on the rules and institutions that govern firms' and consumers' access to data.

Sally Broughton Micova is a CERRE Research Fellow and a Lecturer in Communications Policy and Politics at the University of East Anglia, as well as a member of its Centre for Competition Policy. She is also a Visiting Fellow at the London School of Economics and Political Science (LSE), and a Visiting Lecturer at the Institute of Communication Studies in Skopje, Macedonia. She completed a PhD in Media and Communications at the LSE in 2013. Before entering academia in 2009, she spent over a decade working in international organisations, and continues to serve as an occasional expert for the Council of Europe, EU institutions, and the Organisation for Security and Cooperation in Europe. Her research focuses on policy and regulation in media and communications. Her most recent publications include work on audiovisual media and online platform policy, public service media, and minority language media.

Executive summary

This report analyses the central role of data as an input for the business models that shape competition and innovation in digital markets. By reviewing current data collection practices and the technical processes that transform data into business value, the report sheds light on the economic impact of data in the three cases studies i) online search, ii) e-commerce and iii) media platforms. Based on the insights from these use cases of data, the report considers several policy proposals for data access remedies devised to safeguard competition, innovation and the openness of the digital ecosystem, especially for new entrants. In this context, the report discusses the harms and benefits of data aggregation, and the goal of *digital markets contestability* through improved data access for third-parties. It also highlights the economic trade-offs that policy makers face when considering data access remedies to promote competition and innovation in the digital space.

I. Case studies on online search, e-commerce and media platforms

In the first part of the report, we review three highly popular services in digital markets. In each case study, we analyse the collection and use of data and highlight the economic benefits and competitive advantages that can be derived from data. In particular, we show that in all case studies the ***breadth* and *depth of data* are important determinants of the quality improvements and economic value** that can be derived from data. A broader data set means that information on more users is available, i.e., the data set is more representative and contains on average more data per item. In contrast, a deeper data set refers to the length of the user profiles, i.e., on average there is more data available on each user.

i) Online Search

The first case study on online search highlights how search engines rely on the collection and processing of data resources to retrieve relevant information from the distributed content and documents of the World Wide Web. Here, data plays a key role in improving the quality of search results, which is mainly determined by the ranking decisions of a search engine.

The ***search index data*** that is collected by crawling the publicly accessible web content represents the basis for the matching of users' search queries to relevant websites. Completeness and freshness of the web index data determine the set of search results that can be retrieved by a search engine.

Search query data provides the basis for improving search quality by understanding users' search intent and providing them with suitable results. **Combined with observed *data on user behaviour,* search query logs are used to analyse the implicit quality feedback** given by users' observed decisions and actions on the search engine platform. Collection and analysis of this data is used to improve the ranking decisions and the matching of search queries. Moreover, **personally identifiable *individual user data*** can be used to infer the context of a search request and thus to improve search quality **by personalising the ranking of search results**. User profiles may be created from observing user behaviour on the search engine website, but also from ***tracking user activity*** on other services and in other domains. Inferred information from this data can further improve the quality of search results, especially concerning new search queries. Additional ranking information may be retrieved by analysing the similarity of user profiles and their past behaviour. Collection of ***geographic tracking data*** extends the depth of user profiles by including information on user behaviour in physical environments, which is especially relevant for the quality of local search results.

Ranking criteria based on the accuracy and completeness of third-party-contributed information may create incentives for businesses to **create *third-party business data* directly on the search platform**. Besides, search engines as intermediaries may collect data on interactions and transactions carried out between users and businesses on their platform. This data can be used to improve search quality, but also in other markets where the search engine is active. Moreover, search query data and data on user behaviour can increase the effectiveness of ***search advertising*** and tracking advertising effectiveness on third-party websites can give a search engine access to additional data on user behaviour.

ii) E-commerce

In the second case study on e-commerce, we analyse the role of data for (i) ***demand forecasting***, which is used by retailers to curate product portfolios, for efficient operations and logistics, and for (ii) ***recommendation systems***, which personalise a consumer's shopping experience and facilitate consumers' discovery of new and suitable products in e-commerce markets.

Aggregated sales data **serves as the main input for demand forecasting**, which allows retailers to develop their product portfolio according to observed consumer tastes and also to save costs by promoting efficient logistics, optimal warehousing and automated order systems. Better forecasting performance may be achieved, when time-series data on related products' purchase histories are available and predictions can be based on longer time-series of product sales data.

As operators of ***digital marketplaces***, online retailers are in a special position to **observe *data on third-party businesses***, especially behavioural data on user-to-business interactions and purchasing transactions from these businesses. Based on this data, the efficiency of the overall marketplace can be improved, but the access to this data may also give the marketplace operator a competitive advantage in situations, where it competes directly with these third-party businesses.

Large product catalogues with numerous items per product category, the nuanced differentiation between items, and the availability of a wide set of niche items render ***product discovery*** a major task for online retailers to convert shoppers into actual buyers. To provide users with automated product recommendations, ***data on the user base and the product catalogue*** are necessary inputs. To derive **personalised recommendations** that accurately reflect individuals' interests and preferences, state-of-the-art recommendation algorithms rely on both ***explicit feedback data*** in the form of volunteered product ratings and ***implicit feedback data*** in the form of observed user behaviour.

To overcome the **cold-start problem** of recommendations systems, a minimum amount of feedback data on each user and product are required. ***Data on product characteristics*** and ***user attributes*** can help to mitigate the cold-start problem, but contain complementary rather than substitute information to behavioural user feedback data. Therefore, the continuous collection of feedback data is central to gradually improve recommendation performance. **As a by-product of user behaviour, *implicit fine-granular feedback data* is collected** continuously and at relatively low cost by retailers that already serve an active customer base. **Recommendation accuracy** does not only depend on the scale of data collection, but also on data quality, especially the level of granularity at which data can be observed and collected. Moreover, ***cross-domain data on user behaviour*** from other services can be used to infer more general preference patterns of individual users and to identify new similarity relationships across users. Concerning the role of data for **personalised recommendations**, it is important to recognise that additional user feedback data does not only improve the quality of recommendations derived for the respective individual whose data is collected but also exerts a ***positive externality*** **on the accuracy of recommendations for other users**. This is because a deeper user profile allows for better matches when searching for similar users in the process of deriving a recommendation for another user. This positive externality of additional user data can give rise to ***data-driven network effects***.

The search for new user data and the need for deeper user profiles may incentivise large online retailers to enter new markets. Based on the access to online retailing data resources, in combination with well-developed computational infrastructure and technical expertise, data-rich e-commerce incumbents may indeed be in an advantageous position to **enter other existing or emerging markets**.

In e-commerce, **data-driven quality is not the single dimension along which firms compete** for consumers. On the one hand, the quality of physical products and the respective product design are important competition factors that do not require data inputs. On the other hand, retailers compete for consumers in the price dimension and entrants may poach customers from incumbents by undercutting incumbent providers. This may not be a viable strategy in the long run if it implies perpetual financial losses, but it indicates that there are alternatives to competition in data. Notwithstanding, the case study highlights that **data indeed plays an important role in establishing competitive advantages within e-commerce markets**, and this advantage is likely to grow with access to more data. Thus, data can indeed raise entry barriers for new competitors.

iii) Media platforms

Among media platforms, which we define as those whose service is based on the delivery of content to users and that have some level of responsibility for that content, there are **four main business models**: public service media, subscription, advertising supported and freemium. These are not hard categories, some public media is partly advertising funded for instance, however all collect and use data. As advertising-dependent media platforms compete fiercely with each other for advertising expenditure, they are also competing with media platforms based on other business models for the attention of users. We identify **two main purposes in the collection and use of data**: (i) capturing and retaining users, or in other words ***contributing to the appeal of the platform***, and (ii) ***selling advertising inventory***.

Maintaining appeal centres on ***personalisation*** **and** ***service improvement***. Identifiable personal data is combined with insight from a breadth of aggregate data and with non-personal data on content for personalisation. Improving service can be about interfaces and functionalities and also about improving content choice and/or organisation or even informing content production. One element of service quality is the level and nature of consumer protection, such as from illegal or harmful content, which depends on content data and volunteered data from users.

Though untargeted advertising exists, most advertising on media platforms belongs to one of the **three main types of targeted advertising**, each of which is highly data intensive. ***Contextual advertising*** has become highly sophisticated and can involve deep non-personal data on content. It also involves a certain amount of pseudonymised personal data linked to each campaign used to verify impressions and measure the effectiveness of ads. The other two types primarily use data to predict the potential effect of advertising. ***Segment-based advertising*** relies on the insight generated from broad pseudonymised or anonymised and aggregated data from a variety of sources to create audience segments and then uses deep personal data to identify users belonging to those segments at the ad serving level. In ***behavioural advertising***, prediction and thus targeting is based on detailed user profiles drawing on deep data from the observation of identifiable individual users and inferences about them.

There is a privacy motivated push-back on ***user tracking*****, especially through third-party cookies and fingerprinting. Access to (and consent to use)** ***first-party data*** **is a valuable asset**. Tools for trading advertising on the open web will likely be replaced by in-ecosystem tools that use first-party data. Media platforms that do not have large ecosystems generating depth and breadth of personal data with the accompanied consent are under pressure to ***enable consent to third-parties*** used by advertisers for their inventory to be recognised by demand side tools.

Observed behavioural data is aggregated to feed into metrics that ***measure the effectiveness of campaigns***, such as basic impressions that indicate reach, click through rates (CTR), conversion rates (CVR), and other post-exposure behaviour metrics are tracked for each campaign. For media platforms the ability to collect and use this kind of data on user interaction with the advertising they carry is crucial for establishing the value and demonstrating the efficacy of their inventory. The trade in advertising on media platforms generates *a* ***breadth of non-personal transaction data***, especially when it involves real-time bidding. This data is then used to inform future bidding strategies of demand side actors. When they have access to it, it can also inform the selling choices of the media platforms supplying inventory, such as in setting floor prices. In longer term planning and prioritising advertising business **depends on a** ***continual flow of data into advertiser key performance indicators*** **(KPIs),** which are largely derived from the integration of ***campaign and transaction data***, so non-personal and aggregated personal observed and inferred data. Not having access to continual streams of this data can disadvantage some existing media platforms and may give rise to a cold-start problem for new media and content offerings.

II. The economic value of data

As illustrated across the three case studies, data is at the core of digital services today. For all markets surveyed, we conclude that ***more data, especially more data on user behaviour, will gradually improve the quality of the digital service, albeit at a decreasing marginal rate,*** and allow the firms to generate higher economic benefits along various business value dimensions. This positive feedback loop is what characterizes data-driven markets and leads to ***data-driven network effects*** **that create high entry barriers** for firms that do not have access to such data. Although in all three markets it is feasible to enter with a basic service that does not use (behavioural) data, such a service would often be insufficient to attract users and to grow a viable customer base.

Concerning scale and quality advantages, the considered case studies demonstrate that **data is often created as a by-product of consumers' usage of a service**. The scale of operations therefore directly increases the ***breadth of data*** that is available to a firm. We show that empirical investigations point **to positive but diminishing returns from broader data sets**. When collected data can be associated with individual users, this increases the ***depth of data***, i.e., the average length of a user profile increases and more information per user becomes available. Longer user profiles may play an important role with regard to the economic benefits from increasing data scale. On the one hand, additional user information may yield direct improvements with respect to the performance of algorithms, although marginal benefits diminish with larger depth. On the other hand, **additional user data may reinforce the benefits of broader data sets**. This is, because user data does not only benefit the performance of algorithmic tasks targeted at this individual user but also improves the performance of tasks targeted at other users that are identified as similar users, based on the individual-level data. This may give rise to data-driven network effects even in the absence of increasing returns to scale.

Next to the scale of data sets, the ***quality of data*** **significantly influences the economic value** of data that can be extracted. Moreover, quality requirements will determine the competitive ramifications if firms have unequal access to data. Specifically, the ***timeliness of data*** is important to consider, as consumers' preferences change over time and new relevant items such as products or websites appear in the respective business context. In cases where data outdates quickly, the incumbency advantage of directly observing user behaviour will be especially relevant.

Finally, we highlight that the analysis of data-driven competitive advantages must consider the ***complementary inputs*** that are required for the collection and processing of data. In particular, this comprises **computing and storage infrastructure, skilled human resources and algorithms**.

III. A data-driven theory of harm

We then assess and clarify the underlying theory of harm for data aggregation and data exclusiveness. At its root is the presence of **data-driven network effects**, which likely leads to the tipping of a market, such that only one dominant provider prevails, and which creates high entry barriers. In a **tipped market**, **innovation incentives** of both the incumbent and potential entrants are likely to be lower than in a competitive market. Moreover, data-driven network effects also give rise to a ***domino-effect***, which allows data rich incumbents to enter into adjacent markets, thereby increasing their ability to collect data even more. This is facilitated by **envelopment strategies**, whereby existing services are bundled with the new service.

Particularly, ***ancillary data services*****, such as digital identity management services, web analytics services, or financial transaction services may be viewed with scepticism**, because they allow the collection of even more data across otherwise unaffiliated third-party services. However, in this case, providers of such ancillary data services are not competing and innovating in these markets themselves. Additional harms concerning data access may arise in the context of ***vertical relationships***, e.g., when firms are providing both a platform and act as a provider on the platform. Finally, there is also increasing evidence that data-driven network effects and associated entry barriers harm **venture capital for innovative start-ups** that seek to contest the business model of data-rich incumbents. The reason is that such start-ups often find themselves in a 'kill zone', where they are driven out of the market, either through the incumbent's lower marginal costs of innovation (caused by data-driven network effects) or through acquisition.

Data-driven network effects also bear **inherent *efficiencies*** that must be considered before any policy intervention. Realizing economies of scale and scope in data aggregation, which create entry barriers on the one hand, also generally benefit consumers on the other hand, because they allow to identify and develop products and services that cater to a consumer's individual needs and preferences and create efficiencies that would not have been able otherwise.

IV. Policy objectives: contestability, essential data and niche entry

We argue that **contestability in the narrow sense**, i.e., replacing the incumbent by a more efficient entrant in a process of 'creative destruction', is neither a realistic nor necessarily a desirable policy objective. Even if access to (user) data is facilitated through policy interventions, significant data advantages will remain with the incumbent, not the least because deep personal data is not

sharable without a user's consent. Hence, we suggest that policy makers should focus on **enabling *niche entry and niche growth* and a *level playing field* for competitors in new and emerging markets**.

In this context, we suggest that the discussion of 'essential data' may be futile because **'essential data' in the meaning of the essential facilities doctrine often does not exist**. Market entry is possible without access to proprietary behavioural user data and can be based purely on publicly or otherwise commercially available data. However, in practice ***access to such behavioural data* would be necessary for many instances to offer a competitive service or to develop data-driven innovations in other domains**.

V. Data remedies limiting the collection of user data

We review different possible data remedies that aim at limiting the collection of user data with respect to their technical feasibility and the economic trade-offs involved. These remedies include:

- Data siloing (i.e., preventing aggregation of data originating from different services),
- Shorter data retention periods,
- Prohibiting incumbents from buying into default settings,
- Line of business restrictions, and
- Privacy enhancing technologies.

The general problem with these sets of remedies is that they seek to **achieve a more level playing field in the digital economy by breaking the data-driven network effects** of data-rich incumbents. This diminishes the efficiency of the incumbent and thus also diminishes the ability to create value from data more generally. From a mere economic perspective, we argue that many of these remedies would not be effective in fostering competition and entry in digital markets, although data minimisation may have value in its own right from a privacy perspective.

Line of business restrictions, including vertical separation, may be considered by policy makers under very specific conditions, and as a remedy of **last resort** if data sharing remedies should prove to be ineffective. In particular, we suggest that policy makers should consider the possibility to ***restrict the use of ancillary data services by incumbents***, in so far as they allow to track user behaviour across the entire Internet, e.g. identity management services, financial services or web analytics services. Such services make it very difficult for consumers to truly control to which firm they are providing their user behaviour data, and they undermine exclusive data advantages of niche competitors, which may help them to grow and scale. Moreover, such ancillary data services may often be similarly provided by independent third parties, and with relatively little, if any, efficiency losses.

Finally, ***privacy enhancing technologies* should generally be part of the regulatory toolkit**, but must be tailored to the specific use case and must generally be accompanied with other remedies.

VI. Data remedies facilitating access to broad user data through bulk-sharing

We further consider the application and scope of data sharing remedies that aim at providing access to broad user data. We suggest that, to preserve innovation incentives, **only raw user data (observed and volunteered) may have to be shared**. Moreover, only data that was created as a **by-product of consumers' usage** of a dominant service should be within the scope of mandated data sharing (e.g., search queries or location data); but **not (volunteered) user data that represents the essence of the service itself** (e.g., posts on a social media site). The line may be sometimes difficult to draw in practice, but it is important to make this distinction because otherwise legitimate business models may be destroyed and innovation incentives can be unduly harmed.

Shared data should generally be made available through **standardised interfaces (APIs) in real-time** and continuously.
The most challenging part will be to **balance privacy concerns with maintaining enough level of detail in the data**, such that it is valuable for data-driven innovations by third-parties. We survey several **technical and institutional means** that can facilitate this balancing act and prevent de-anonymisation of shared data sets. Within limitations, we entertain the idea that a ***data trust*** **and** ***data sandboxing*** **(at a data trust)** may be feasible if confined to subsets of the data to be shared, particularly with a focus on recency, and if confined to a few select algorithms that may be trained at any given time. The EuroHPC, a European collective effort to create a supercomputing ecosystem, may be the technical host to such a data trust. Furthermore, we see some merit in the proposal to **declare deliberate de-anonymisation efforts illegal** under European law.

VII. Data access remedies in online search, e-commerce and social media platforms

We make specific proposals to advance the debate on broad user data sharing in the context of our three case studies. **Concerning** ***search***, we suggest three categories of data from which data access requests should be considered**:** ***Data on the search query***, ***data on the search results page***, **and** ***data on the user***. Generally, complex trade-offs are to be considered and we suggest that mandated access to data needs to be done **on a** ***case-by-case basis*** **and requires a** ***vetting procedure*** of the data access seeker by the regulatory authority. This will likely come alongside with additional responsibilities and safeguards for the data recipient. At the same time, a **less detailed, highly anonymised data set should be made** ***publicly available*** **without prior vetting**.

Concerning ***e-commerce***, we are sceptical that any mandated sharing of broad user data would be warranted, albeit the ***transparency of data use*** **as well as the** ***detail and mobility of information*** **that is already provided by platforms** could be improved. Competition in and for e-commerce markets is already intense, and not only focused on data use but also on price. Also given the increased e-commerce related activities of data-rich incumbents from other markets, regulatory forbearance for mandated data sharing seems to be for the time being.

In the context of ***advertisement-supported social media platforms***, there may be specific cases of vertical integration or competition concerns relating to ancillary markets, where line of business restrictions might be called upon as a last resort. However, we suggest relying on data sharing remedies aimed at ensuring **continual access to the data necessary to compete effectively** in the first instance. Here, the most contentious issues relate to access to certain categories of aggregate **campaign data** and **user interaction with advertisements**. The sharing of any identifiable personal data, even observed data from the use of the platform or exposure to advertising, is justifiably limited by data protection rules. However, if consumers were allowed to opt into the sharing of their usage data with individual content creators, including a unique identifier, when they consume content on the platform privacy concerns could be alleviated. A case can also be made for levelling the playing field through the **sharing of aggregate performance data** at the level of independently audited audience measurement accessible by all industry participants.

VIII. Data remedies facilitating access to deep user data through continuous data portability

Finally, we discuss how **access to *'deep' raw user data* can be facilitated by strengthening consumer rights above and beyond their existing data portability right under Article 20 GDPR**. In particular, we suggest that in several cases competition and innovation would benefit if firms were obliged to provide consumers with the possibility to **consent to *continuous, real-time data portability***. The scope of data to be transferred should be identical as under Article 20 GDPR. However, to date more legal certainty is needed for the precise scope of Article 20 GDPR with respect to observed (user behaviour) data. Generally, as in the case of mandated sharing of broad user data, **only raw user data (volunteered and observed) should be subject to data portability**. Also, consumers must need to **consent** to every such continuous transfer. Continuous data portability should be made possible through **standardised APIs**, enabling both business-to-business data transfers, but also the use of Personal Information Management Systems (PIMS). Demonstration projects like the Data Transfer Project and Solid exemplify that such continuous data portability is feasible from a technical perspective. However, mandating continuous data portability will require policy makers also to facilitate the setting of and agreeing on (open and secure) **standards for data transfers, and consumer consent**.

01

INTRODUCTION

1 Introduction

This report analyses the central role of data as an input good for the business models that shape competition and innovation in today's digital markets. By reviewing current data collection practices and the technical processes that transform data into business value, the report attempts to shed light on the economic impact of data beyond an often-cited, but an abstract notion of big data value. Based on a more nuanced understanding of the economic and technical properties of data and its use across selected market settings, the report investigates recent policy proposals devised to mitigate the market power of data-rich incumbents and to safeguard the openness of the digital ecosystem for new entrants. In this context, the report discusses the **goal of *digital markets contestability* and highlights the economic trade-offs** that policy makers face when considering data access remedies to promote competition and innovation in the digital space.

Recent reports on competition in digital markets have emphasised that to avoid long-term monopolisation, it is indeed vital to **ensure that digital markets remain contestable**. This conclusion has been drawn in the context of growing concerns about increasingly concentrated digital markets, where a few data-rich firms have gained prominent positions and large user bases across horizontal and vertical markets. These concerns about contestability are also shared by policy makers. For example, the European Commission (2020a[1]) has stated that "many online businesses have struggled with systematic problems familiar to the platform economy regarding contestability, fairness and the possibility of market entry" when it announced its consultation on the Digital Services Act.

Besides network effects and digital platform business models, policy debates on digital markets have frequently referred to data and data-driven business models as potential impediments to the contestability of these markets. In its European strategy for data, the European Commission (2020b[2]) has emphasized that „a small number of Big Tech firms hold a large part of the world's data" and then concluded that "[t]his could reduce the incentives for data-driven businesses to emerge, grow and innovate in the EU, today" (p. 3). In this spirit, policy reports have suggested to "advance data openness where access to non-personal or anonymised data will tackle the key barrier to entry in a digital market, while protecting privacy" (Furman et al., 2019[3], p.6).

In this report on digital markets contestability, we focus specifically on **the implications of data access and data use on competition and innovation**. This is not to say that other characteristics of these markets such as network effects are less important, but they have been discussed extensively and in detail in several recent policy analyses (see, among others, Crémer et al., 2019[4], Furman et al., 2019). Although most of these analyses also consider data-related issues, the discussion of data use cases and data's role for economic value creation often remains rather general given the wider scope of the reports. However, as Crémer et al. (2019) state, "[d]iscussing access to data in the abstract is futile" (p.73). Therefore, we attempt to complement these reports by evaluating data-related policy goals and data access remedies based on an economic analysis grounded in the review of selected case studies well as recent empirical and theoretical findings of the academic literature.

To this end, our analysis focuses exclusively **on policy remedies that are directly related to access to data**. This includes remedies that aim to limit the data access for specific market participants as well as remedies that are targeted to advance data openness and facilitate data sharing. However, we do not consider alternative or additional non-data related remedies that could be devised to mitigate concerns about competition or innovation issues that may arise from an entrenched dominant position of a data-rich incumbent. Such remedies may indeed be suitable alternatives to the data access remedies discussed in this report, but they are outside of the scope of our analysis.

---

[1] European Commission. (2020). The Digital Services Act package. Available at https://ec.europa.eu/digital-single-market/en/digital-services-act-package

[2] European Commission. (2020). A European strategy for data. Available at https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

[3] Furman, J., Coyle, D., Fletcher, A., McAules, D., & Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. *Report prepared for the Government of the United Kingdom, March*.

[4] Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). Competition policy for the digital era. *Report for the European Commission*.

Moreover, **in this report, we are especially concerned with the role of user data** due to its prevalent use in today's most popular digital business models and its widespread collection by firms in the digital space. In the case studies, we discuss the role of some non-personal data, such as data stemming from content and product categorising and indexing or from financial transactions. However, where firms have exclusive access to non-personal data, such data resources may be qualified as essential if competition concerns arise and data access remedies may thus be considered and devised under European competition law and Article 102 TFEU. This requires a very high threshold to be met and is not unproblematic,[5] however, the personal data of users poses particular challenges and therefore our consideration of potential remedies focus primarily on this type of data.

Concerning user data, **competition law will often not apply or will need to be complemented by ex-ante regulation due to the complex and dynamic nature of digital markets**. Thus, the policy framework under which access to user data can be deliberated is a priori less clear. Hence, this report lays out the potential theory of harm in the context of user data and discusses the policy objectives that can be achieved by data access remedies. In consequence, we focus on **ex-ante regulatory remedies that can promote entry and growth in data-driven markets**. In our evaluation of policy interventions, we analyse the economic trade-offs involved in the alternative data remedies that may be applied to either limit the data access of incumbents or to facilitate the data access for competitors. In contrast, this report does not consider the governance framework and the respective institutions that are necessary to implement the various data remedies. This will be covered explicitly in the companion CERRE report by Feasey and de Streel (2020).

Finally, it is important to highlight that this report is **foremost concerned with an economic analysis** of the reviewed use cases and the involved trade-offs of the various policy proposals. Besides, when feasible and instructive, we consider technical aspects of the use of data in view of implementing specific policy proposals. However, we do not attempt to provide a holistic and exhaustive analysis of the ex-ante regulation of digital markets and platform markets in particular. This would require including, inter alia, plurality concerns, a wider citizens' rights perspective or considerations of fairness and economic well-being. However, we occasionally refer to some of these considerations in the respective policy contexts.

Structure of this report

To evaluate the role of data in today's digital markets, Section 2 starts by investigating the **role of data for service quality and competition in three key digital markets**. Firstly, we investigate the data resources that are collected and used in the domain of *online search*. Secondly, we describe the collection and processing of data in *e-commerce* with respect to demand forecasting, ancillary e-commerce services for third-parties and personalised recommendations. Thirdly, we examine the role of data for digital *media platforms*. Next to the collection and use of data for personalisation, service improvement and algorithmic content moderation, data is especially relevant for the sale of targeted advertising in the case of advertising-financed media platforms.

In Section 3, we then draw on the specific insights from the case studies and the academic literature to characterise in more general terms the criteria that determine the **economic value and competitive advantages from data**. We scrutinise (i) whether and to which degree other firms can duplicate data resources of data-rich incumbents and (ii) how the economic value from data is related to the breadth and depth as well the quality of data. Finally, we highlight possible data-driven network effects that promote concentration within digital markets and also facilitate the entry of data-rich firms into other markets.

Based on the evidence gathered across the case studies and the economic characteristics of data, we then, in Section 4, assess the **data-driven theory of harm** that would warrant policy interventions in the form of ex-ante regulation and discuss the primary **policy objectives of data access remedies.** In particular, we contrast the concept of contestability with the broader goal of niche entry and growth and conjecture what can ultimately be achieved by the means of data remedies.
In Section 5, we turn to **the evaluation of possible data access remedies**. At first, we analyse *data* access *remedies that would limit the collection of user data by already data-rich incumbents* to

[5] For discussion see: Inge Graef, *EU Competition Law, Data Protection and Online Platforms: Data as Essential Facility: Data as Essential Facility* (Wolters Kluwer, 2016); See also the companion CERRE report on Digital Markets Contestability by Feasey and de Streel (2020).

level the playing field for competitors and new entrants. We then highlight alternative policy interventions that would facilitate the sharing of user data by opening up access to raw data resources collected by data-rich incumbents. In this context, we propose a dual approach with two complementary types of data access remedies. One set of remedies may aim at *facilitating access to 'broad', anonymised raw user data through bulk sharing*, whereas other remedies may be designed to *facilitate access to 'deep' raw data that contains personally identifiable information through continuous data portability*.

Finally, Section 6 concludes by **summarising the insights** of the report and pointing to specific governance and implementation issues that arise from our economic analysis of data remedies.

The terminology used in this report

Before reviewing the role of data in the different case studies on digital business models, it is helpful to introduce and define some of the terminologies that we use throughout this report. As defined in the European General Data Protection Regulation, we will refer to **personal data** as "any information relating to an identified or identifiable natural person"[6]. Even if data is not personally identifiable, data may contain unique identifiers, which allow the combination of data points that relate to the same individual. We will call this property *traceability* and refer to the respective data as *pseudonymised data.* In practice, the collection of pseudonymised data is especially relevant, as web users are frequently recognized and traced by the means of browser cookies, which are sent as part of a regular website request to the respective website owner.

Furthermore, for data collected from users, we will differentiate between the following types of data:

- ***Volunteered data*** is explicitly and intentionally revealed by a user, such as a name and a birthday entered into a registration form, a post, tweet or rating submitted, or an image or video uploaded. Consumers are usually aware of the volunteered data that they reveal and often this is the only type of data that consumers think they have revealed when using an online service.

- ***Observed data*** is obtained from the usage of a device, website or service and the user may or may not be aware that such data is collected. This ranges from clicks on products and purchase histories over geo-locations gathered by GPS sensors in smartphones to recording every single interaction of the consumer with the service – potentially even when the consumer does not even know that she is currently interacting, such as in the context of voice assistants that are constantly recording.

- ***Inferred data*** is derived through refinement and recombination from volunteered and observed data, e.g. by use of data analytics such as clustering, filtering or prediction. The result can be a complex preference profile of a consumer or a recommendation. Inferred data can already be the *knowledge* that in turn can provide actionable insights. Thus, inferred data is ultimately the basis for competition between data-intensive firms, whereas volunteered data and observed data are the 'raw data' inputs.

Concerning the size and the dimensions of data sets, we will frequently refer to the *breadth* and *depth of data*. A **broader data set** then means that information on more users is available, i.e., the data set is more representative and contains on average more data per item. In contrast, a **deeper data set** refers to the length of the user profiles, i.e., on average there is more data available on each user. Both dimensions and their implications for the value of data are discussed in more detail in Section 3.

[6] Article 4 2016/679

02

CASE STUDIES

2 Case Studies

In this section, we review three business models that provide highly popular services in digital markets: online search, e-commerce and online media. For these business models, we analyse the collection and use of data and highlight how data is used to create economic benefits and competitive advantages.

2.1 Online search

Online search is the main gateway to the content of the World Wide Web. The most popular search engine Google Search now handles over 80,000 search queries per second.[7] The key objective for search engine operators is to retrieve the most relevant information and websites in response to each of these user-entered search queries. To this end, the collection and analysis of data are at the core of the search engine business model. Beyond the retrieval of websites, general web search enables users to also search for and directly access other media content from the web such as pictures and videos. As mobile devices have become ubiquitous, traffic from mobile search has overtaken web search from stationary devices.[8] Thus geographic location information and knowledge about user movements are increasingly important for finding relevant search results in proximity to the user. Local search, which identifies relevant physical entities such as stores and local businesses in response to a search query, has been a particular focus of recent search innovations.

In the following, we review the basic architecture of a search engine and the main factors that determine the quality of a web search engine. Along with this review, we focus on the components and quality factors that involve the collection and use of data. We highlight the different types of data that are collected and discuss their technical and economic value for the search engine business model. We also discuss data-related issues in the context of local search and search advertising.

2.1.1 Search engine architecture

Search engines consist of three basic building blocks: the search index, the query engine and the search interface. These components support the two main tasks of search engines, the *indexing process* and the *query process* (see Croft et al., 2015[9]). The indexing of searchable content is undertaken before users interact with the search engine. The querying process then encompasses the execution of user queries and the look-up, ranking, and display of relevant results from the search index. Before we evaluate the major quality criteria for web search services, we briefly describe the three main components of an online search engine and the associated functions.[10]

2.1.1.1 Search index

Search queries are executed against the search index, which is an organized offline copy of the entire web content. To collect web documents and to keep an up-to-date copy of these documents, search engines are frequently crawling the web. Algorithmic crawlers start from a specified seed set of sites and download the websites that they visit. Downloaded sites are then parsed to find link tags that refer to the addresses of new websites, which are subsequently crawled (the "frontier"). To avoid stale copies of websites in the index, crawlers must continually revisit the contained websites. Usually, crawling and updating of the index are prioritised based on the relevance or importance of a website. Due to the size of today's World Wide Web, the search index amounts to a size of well over 100.000 Terabyte[11], which must be stored by the search engine operator. This is especially challenging because the processing of search queries and the look-up of index entries must be executed in real-time and without noticeable delay for users.[12] Therefore, Google, for example, keeps its index in memory across distributed servers rather than stored on hard disk drives.[13] This requires efficient data structures and network technology as well as large investments in hardware infrastructure. For example, Google has developed its own proprietary data structure, Google BigTable, to meet the requirements with respect to large-scale and real-time operations.[14]

---

[7] Internet Live Stats, 2020, https://www.internetlivestats.com/one-second/#google-band
[8] https://www.thinkwithgoogle.com/consumer-insights/mobile-search-trends-consumers-to-stores/
[9] Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search engines: Information retrieval in Practice*. Reading: Addison-Wesley.
[10] The descriptions of search engine functions in this Section are mainly based on Croft et al. (2015).
[11] https://www.google.com/search/howsearchworks/crawling-indexing/
[12] https://www.thinkwithgoogle.com/marketing-resources/the-google-gospel-of-speed-urs-hoelzle/
[13] http://glinden.blogspot.com/2009/02/jeff-dean-keynote-at-wsdm-2009.html
[14] https://cloudplatform.googleblog.com/2015/05/introducing-Google-Cloud-Bigtable.html

To improve search efficiency, the content of downloaded websites is further processed before it is added to the search index. The main objective is to convert the content of websites into index terms to which the search queries than can be matched. Specifically, this entails text processing, such as parsing, stopping and stemming of words contained in the websites in order to convert different versions of a word into a more consistent index term. Furthermore, the structure of webpages is analysed. Link analysis exploits structural HTML code tags to extract important words and outgoing links to other websites. Moreover, information extraction techniques may identify classes of words that are particularly relevant such as names of people and organizations. For index creation the collection of document statistics is central. The counts and positions of index words provide an important basis for ranking algorithms when queries are matched based on user-entered or algorithm-inferred keywords. Furthermore, index terms are frequently weighted such that words that convey more specific information receive higher weights than words that are rather generic. Finally, the index is inverted such that the document can now be identified by its index terms.

Concerning the search index, the goal is to achieve high coverage (completeness of indexing) and freshness (up-to-date indexing). Next to keywords as index terms, search engine operators store additional quality signals in the index. For instance, the Google Search Index contains key signals that the search engine considers for its ranking "from keywords to website freshness" of hundreds of billions of webpages.[15] With the technology update Caffeine, introduced in 2010, Google started continuous crawling and incremental updating of its search index, whereas before the entire index had to be replaced in order to update its entries.[16] Thus, Google can detect web changes almost immediately and include this updated information in the search results.[17] Besides, Google stores semantic relations between more than one billion real-world entities (people, places and things) in its Google Knowledge Graph.[18] This semantic information is used to provide users with immediate answers on the search result page for about a third of all search queries according to reports in 2016.[19]

Search index data: Data on the content and websites of the World Wide Web is the core input for web search engines. Data collection is done through web crawling of publicly available information and determines the completeness and freshness of the index data.

2.1.1.2 Query engine

Given the search query of a user, the task of the query engine is to retrieve the most relevant results from the search index. Because keywords in the search query are often poor descriptions of the actual information need of a user, additional computational processing of the query is required to increase the quality of search results as perceived by the user. The context of a query as well as user interaction is especially important for understanding true user intent. In general, the query process encompasses the following steps:

a) *User interaction*: Search queries are received as input from the search interface and further transformed before being matched against the search index. Queries are spell-checked, further refined or expanded in order to find relevant matches. Transformed queries may be executed internally or suggested as explicit feedback to the user.

b) *Results output*: Search results are usually presented as short text snippets and with highlighted words, so users can evaluate whether a website matches their actual interest. Furthermore, search engines may display different types of results in different formats. For example, a search engine may present users information from websites or answers to the query directly on its search result page if it associates a search query with a specific information need. In addition to the display of organic search results, search engines regularly display advertisements as sponsored search results.

---

[15] https://www.google.com/search/howsearchworks/crawling-indexing/
[16] https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html
[17] http://glinden.blogspot.com/2009/02/jeff-dean-keynote-at-wsdm-2009.html
[18] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges. *Queue*, *17*(2), 48-75.
[19] https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/

c) *Ranking*: Due to the tremendous number of websites, the ranking of search results largely determines search quality and is thus among the core tasks of today's web search engines. Only by selecting and sorting search results according to their relevance for the inferred search intent can information overload of the human user be avoided. Traditionally, search rankings were mainly based on how often keywords from the search query were contained in a web document. However, today, many more ranking signals are combined to determine the ranking score of a website and the relative order of display on the search result page.

d) *Evaluation*: To analyse and improve the perceived quality of displayed search result pages, search engine operators track user behaviour on their website. Individual search queries and user interaction with search results such as the click-through are observed and logged. As will be detailed in the following, this data can then, among other purposes, be used for improving future interactions with consumers (e.g., new query suggestions), as additional or adjusted signals for the ranking algorithm, and the display of relevant advertisements. Moreover, individual tracking data on search behaviour can be used for personalisation of future searches.

2.1.1.3 Search interface

The search interface handles user interactions and provides users with access to the content of the search index. To achieve high user satisfaction, technical performance of the search engine is a particularly important criterion. The speed of information retrieval is viewed as critical to foster user engagement with a search engine. For example, Google estimates that a 400ms delay in delivering search results leads to a 0.44% drop in overall search volume.[20] This requires large supply-side investments with regard to computationally efficient algorithms, high-performance computing hardware and distributed storage capacities.

Recently, personal assistants like Alexa from Amazon or Siri from Apple have popularised a class of new search interfaces that can be queried by the human voice and through a variety of devices. As search results are shortened to few or even to a single search result, ranking decisions become even more important. Furthermore, natural language processing of queries and knowledge about the context of a search and the user become vital inputs for retrieving relevant (web) content. At the same time, voice input allows the search interface to capture additional audio signals, which could be used to infer more context-specific information (such as the sentiment of a search query) from how the query is posed by an individual.

2.1.2 Web search

The quality of a search engine is critical to attract users. Next to technical performance parameters, such as the speed of the search process, quality is critically determined by whether users perceive the search results as relevant for their information need. To this end, the completeness and freshness of the web index play an important role as necessary inputs for relevant search results. However, due to the large scale of web content and the limited time of searchers, the selection and ranking decisions can be viewed as the pivotal criteria that differentiate the quality of search engine operators. To this end, the quality of search results is vital for being able to attract consumers and thus to being competitive as a search engine. In particular, the quality of ranking decisions is therefore seen as a key competitive factor for search engine operators.[21]

The definition of a global relevance metric on which basis the quality of ranking decisions can be measured is non-trivial. Most prominently, Google has established the E-A-T principle, which characterizes expertise, authoritativeness, and trustworthiness of a website as the main evaluation criteria for whether a search result and its associated ranking position are of high quality with regard to a specific search query.[22] To monitor and improve the quality of the Google search algorithm, human quality raters are continually evaluating search results on the basis of these three main criteria. The quality evaluation is then used to adjust the weights of various signals that are taken into account by the ranking algorithm.

---

[20] https://www.thinkwithgoogle.com/marketing-resources/the-google-gospel-of-speed-urs-hoelzle/
[21] Chapelle, O., & Chang, Y. (2011, January). Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge* (pp. 1-24).
[22] https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

Modern ranking algorithms take into account a large number and variety of ranking signals when scoring the relevance of search results for a specific query. For example, Google's ranking algorithm considers over 200 ranking signals based on different data inputs. The relative weights of these signals for the ultimate ranking decision are constantly adjusted and optimised. Google, for instance, is making updates multiple times every day above and beyond major updates of the core ranking algorithm.[23] In 2019, Google made a total of 3620 changes, which amounts to almost 10 updates per day on average.[24] The effectiveness of these changes is evaluated based on i) the feedback of human quality raters, ii) sided-by-side experiments, where quality rates choose between two alternative rankings of results and iii) live traffic experiments (about 50 per day), where samples of representative users are exposed to the changes. The effect of a live-experiment is then evaluated with regard to the query volume, the number of queries that were abandoned, users' clicks and time until clicks occur.[25]

In the following, we summarize the main categories of ranking signals and highlight the data sets that are used by online search engines to determine the ranking of search results. Given that search engine operators themselves optimize these rankings with respect to the overall quality of search, we can infer that these signals and specifically the associated data sources are relevant inputs to offer users a state-of-the-art web search service.

2.1.2.1 Search intent and query understanding

Based on the user-submitted search query and the context of an individual search, search engines attempt to infer the searcher's intent behind the query to identify the keywords to look up in the search index. They may also assist the user in reformulating the original query to find more relevant search results. Moreover, the search engine may classify the type of information request (e.g., specific vs. broad information needs), which can alter how keywords are interpreted and may trigger different result formats in the search results. For example, Google displays information directly on its search result page if it classifies the query to request information about a current event, such as the score of a soccer match.

Whereas, traditionally, search queries were interpreted mainly based on the keywords entered by the user and analysed based on text statistics, modern search engine attempt to incorporate the context of a query and take into account semantic relationships between keywords. For example, Google uses query expansion techniques based on its Synonyms System, which instead of searching only for the keywords entered by the user, also takes into account search results for semantically related keywords.

Based on the analysis of historic search query data and the context of the entire query, the system learns semantic synonyms, which add relevant search results, but also so-called "siblings", which should not be equated with keywords from the original query despite similar meanings. Google engineers have referred to this neural matching system as "one of Google Search's most important ranking components".[26]

More generally, machine learning models have recently been integrated into search engine systems to infer the user intent behind queries based on large data sets that capture user behaviour. Machine learning approaches are specially used for aggregating and weighting various other ranking signals based on the identified intent behind a query.[27] Thus, these approaches may also introduce new composite signals.[28] Whereas initially Google relied exclusively on rule-based algorithms for a long period of its existence, in 2015, the company confirmed that it had integrated the machine-learning algorithm RankBrain into its search engine.[29] Before, there was scepticism towards machine learning techniques in organic search, because rule-based mechanisms could be understood and tweaked more easily.[30] In contrast, deep learning algorithms are often tweaked based on trial-and-error,

[23] https://moz.com/blog/how-often-does-google-update-its-algorithm
[24] https://www.google.com/search/howsearchworks/mission/users/
[25] Ibid.
[26] Haahr, P. (2019). Improving Search Over the Years (WMConf MTV '19). Available at https://www.youtube.com/watch?v=DeW-9fhvkLM&, https://searchengineland.com/googles-neural-matching-versus-rankbrain-how-google-uses-each-in-search-314311
[27] https://www.seroundtable.com/google-explains-machine-learning-search-28697.html
[28] https://searchengineland.com/google-uses-machine-learning-search-algorithms-261158
[29] https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines
[30] https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/

which requires experiments and large sets of training data. One year later, Google announced that RankBrain was now applied to 100% of its search queries and acknowledged it to be among the most important signal of its ranking algorithm.

In contrast to neural matching, which relates keywords to searches, RankBrain is designed to understand how websites are related to semantic concepts.[31] Thus, RankBrain can match search queries to semantically relevant websites even if the websites do not contain literal keyword matches. Understanding semantic concepts are especially valuable for complex and ambiguous queries, conversational queries and long-tail queries that are rarely or never observed before as the algorithm can find more popular keywords that correspond to the same concept. Machine-learning algorithms like RankBrain are trained based on batches of historical search queries, which may also include data on user interaction (see 2.1.2.4).[32] Training is conducted offline, i.e., outside of the live production environment. Thus, implications of model changes can be tested on samples of search queries, before they are deployed to the live system. Learning to match models can complement traditional matching methods of exact query terms and thus improve matching quality of relevant documents.[33]

Current research on neural information retrieval studies learns to rank models that rely simply on the raw text of a search query and a website document as input to yield relevant rankings of search results.[34] Although such models are not conceptually new, their training demands large-scale data sets. While unsupervised learning approaches require only data on search queries and website documents, supervised learning approaches additionally learn from labels that relate relevant websites to a specific search query. Such labels can, for example, be inferred from historic click data of searchers.

Search query data: Analysis of search queries allows a search engine to improve its understanding of search queries and the search intent of a user which can be used for enhanced matching of results to queries.

2.1.2.2 Content quality and keyword fit of a website

The relevance of an individual search result is determined to a large degree by the content of the website. On the one hand, it is pivotal that the content matches the search query and, on the other hand, the content itself should be of high quality in the sense that it contains accurate and trustworthy information. To assess the content of a website, search engines therefore regularly consider

a) *On-site relevance indicators*: The fit of a website's content is primarily determined based on the matching of keywords. If the website contains the same or similar keywords in important positions it will likely be considered to be relevant for the search query. Depending on the inferred intent behind a query, the goal may also be to provide users with new and complementary information to satisfy a user's information need (see Section 2.1.2.1).

b) *Off-site relevance indicators*: General quality of a website's content can be inferred from its relationship with other websites. Incoming links, citations on other websites, the authority of referral sites and the link acquisition rate over time may, among other factors, be analysed and considered as indicators of high content quality. The early success of Google Search has largely been attributed to its PageRank algorithm, which uses the number and quality of links to determine a website's relevance. PageRank is still used as a rating signal, but its algorithmic computation has been made more efficient[35] and its link concept has been refined to also consider topic-specific authority and trust of a website according to a patent filed by Google in 2018.[36]

[31] https://www.rankranger.com/blog/neural-matching-rankbrain-difference
[32] https://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440
[33] Mitra, B., Diaz, F., & Craswell, N. (2017). Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1291-1299).
[34] Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, *13*(1), 1-126.
[35] http://www.seobythesea.com/2018/04/pagerank-updated/
[36] https://patents.google.com/patent/US9165040B1/en

The key data source for analysing the content quality and keyword fit is the search index. For the evaluation of off-site relevance, the search engine may undertake additional crawling and analysis of link structures within the retrieved web documents.

2.1.2.3 Usability of a website

The general quality of a website is also affected by technical parameters and its general usability. Ranking algorithms, therefore, take into account the quality of service (QoS) criteria that measure the usability of a website. These criteria may concern the layout and structure of a website, but also the load time of a website and security features. To incorporate these factors in the ranking algorithm, search engines collect data by testing technical features of websites in the search index and their appearance on different devices and in different browsers. By defining the benchmarks for good usability, a popular search engine like Google has thus significant impact on the design of websites in the global web. For example, the emphasis on the loading speed of mobile pages or the introduction of HTTPs encryption of a website's traffic as a positive signal in its ranking algorithm has led to widespread upgrades and adoption by websites.[37] In this context, Google often provides website owners with its tools to analyse the performance of their websites (see, e.g., Google Lighthouse[38]), which may provide the search engine operator with additional access to data from third-party websites.

2.1.2.4 User behaviour

User behaviour and specifically their interaction with search results may reveal valuable information on the quality of search results as judged by users. Therefore, search engines collect augmented search query logs that in addition to the query term contain user identifiers, search result pages with a list of URLs and ranks of the individual results, user clicks and timestamps (Croft et al., 2015). From this data, it is possible to infer relevance judgements of users when they decide which action to take after entering the search query, although biasing factors such as the rank on the result page must be taken into account. Specifically, click-through data on individual search results can be used to predict preferences between pairs of websites, while aggregate click distributions can be used to identify individual search results that outperform the average performance, e.g., given the rank on the result page (Croft et al., 2015). This implicit feedback has been found to drastically improve web search rankings.[39] Moreover, metrics like a user's time spent on a website before returning to the search engine or the bounce rate of searchers that immediately return after clicking a result enable search engines to infer the relevance of a website with regard to a specific search query.

Hence, information extracted from data on user behaviour can serve as a valuable quality signal for the ranking algorithm, because it integrates the users' revealed preferences and allows for continuous improvement of search engine results. In some cases, users' behaviour may signal the actual information need of a search more effectively than the search query itself, which can be difficult to express for searchers who are unfamiliar with the topic that tries to find out about. Google reports that it uses and processes "aggregated and anonymised interaction data" as an input for their machine-learning systems to evaluate the relevance of search results. Moreover, Google engineers have confirmed that "the ranking itself is affected by the click data".[40]

Continuous data collection also allows search engines to adjust rankings dynamically if users' tastes change over time or if the semantic context of a specific query changes and thus requires a modified ranking. For example, while search queries with the keyword "Corona" have only recently most likely referred to the popular beer brand, this has changed dramatically with the Covid-19 pandemic. Furthermore, according to Google, 15% of its daily search queries are new.[41] Therefore, the collection of data on user behaviour contributes to new information that may be especially valuable for the ranking of search results in response to long-tail queries, i.e., keywords that are only rarely entered.

---

[37] See, e.g., the announcements on considering loading speed of websites https://developers.google.com/web/updates/2018/07/search-ads-speed and on mobile-first indexing https://webmasters.googleblog.com/2020/03/announcing-mobile-first-indexing-for.html

[38] https://developers.google.com/web/tools/lighthouse

[39] See, e.g., Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 19-26).

[40] https://moz.com/beginners-guide-to-seo/how-search-engines-operate

[41] https://blog.google/products/search/our-latest-quality-improvements-search/

Data on user behaviour: Augmented search query logs that record click behaviour on search results pages are the main input for analysing users' implicit quality judgements. Data on user behaviour is thus used to adjust rankings of search results and improve the matching quality of keywords to search results.

2.1.2.5 Personalisation of results

Finally, web online search has evolved from the retrieval of information that is assumed to be of universal relevance to a more personalised approach that considers the context of each search and user. Several data categories are collected by search engines and taken into account to personalise the ranking of search results. First, *geographic location data* from which the search originates is observed from the self-reported address by the user, the IP address of the connected device or the GPS signal of the device. Second, *long-term search history data* of a user that spans across several web sessions can be collected based on authenticated user accounts or through tracking cookies. Besides, *data on short-term search behaviour* in the same session can also be identified by browser information or the IP address and may convey information about the greater context of a user's search intent. Third, *technical context data* such as information on the user's device and the preferred language can be collected based on the configuration parameters sent with browser requests. Fourth and more generally, *data on user activity* and *data on user attributes* may be collected from other online services that the search engine operator provides to users or external services. For example, Google states that it personalises search features based on a user's activity that were connected with its Google account.[42] This allows the integration of data from a wide variety of services in different domains, such as movement profiles from Google Maps or browsing histories from Google Chrome. Similarly, search engines may make use of social network information, such as friend connections or a user's 'likes'[43].

Ranking according to personalised signals can improve the relevance of search results as the context of a specific search can be inferred more accurately.[44] To this end, inferring the demographic context from data on user attributes as well as inferring the situational context from the time and location data of a search request can improve the relevance of search results.[45] Furthermore, the ranking may account for group-level or individual-level preferences that can be inferred from similar behaviour of users on other services or in other domains.[46] Moreover, data on user behaviour on other websites and services can be used to complement data on users' interaction with search results. By augmenting the data on user behaviour, ranking signals may become more accurate as data becomes less sparse.[47] Finally, personalisation may be achieved by expanding the user-entered query in different ways according to user-level information (*query augmentation*).[48] For example, search queries containing acronyms or names of people may be matched with different related keywords depending on the context inferred from data on individual short-term search behaviour.[49] In consequence, this can improve the matching between the original search query and the retrieved search results.

Individual user data: Information on individual users and user behaviour can be used to tailor personalised search results to individual-level contexts and preferences. Data for creating user

[42] https://www.google.com/search/howsearchworks/algorithms/
[43] SearchEngineNews.com (2020). The UnFair Advantage Book Winning the Search Engine Wars. Available at https://searchenginebook.com
[44] https://search.googleblog.com/2011/11/some-thoughts-on-personalisation.html, Tamine, L., & Daoud, M. (2018). Evaluation in contextual information retrieval: Foundations and recent advances within the challenges of context dynamicity and data privacy. *ACM Computing Surveys (CSUR)*, *51*(4), 1-36.
[45] Kharitonov, E., & Serdyukov, P. (2012). Demographic context in web search re-ranking. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2555-2558); Zamani, H., Bendersky, M., Wang, X., & Zhang, M. (2017). Situational context for ranking in personal search. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1531-1540).
[46] See, e.g., Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 449-456).
[47] Matthijs, N., & Radlinski, F. (2011, February). Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 25-34).
[48] Pitkow, J. (2002). Personalized search: A content computer approach may prove a breakthrough in personalized search efficiency. *Communications of the ACM*, *45*(9), 50-55.
[49] Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J. T., Chen, E., & Yang, Q. (2009, July). Context-aware query classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-10).

profiles may be collected from users' activity on other services of the search engine operator and can be combined with data on individual search history.

2.1.3 Local search

As mobile search traffic has been steadily growing and has now surpassed other search traffic by an estimated share of about 60% of total search traffic[50], local search for physical businesses that engage in face-to-face contact with their customers is becoming increasingly important to satisfy consumers' information needs. As search engines serve more and more local information requests, this has important ramifications for physical stores and service-area businesses, as their demand is significantly affected by how high they rank in these search engines. According to surveys by the marketing firm BrightLocal, 90% of US consumers have used the internet to find a local business, with 33% looking every day.[51] According to Google Analytics data from a sample of local businesses, 36% of local businesses' website traffic in 2018 came from mobile sources.[52] Because web search engines represent a natural starting point for users to also begin their search for a local business, they are viewed as important referrers of customers. According to market research by Google, four in five US consumers used search engines to find local information, 30% of all mobile searches are related to location and 76% of smartphone users visited a physical store within a day of their local search.[53]

2.1.3.1 Ranking signals, additional services and third-party data

To account for the specifics of users' search intent in the context of local search, web search engines like Google account for additional ranking signals, especially proximity, and present users additional result displays if they categorize a search query as a local search request. For example, Google may present users with a "Knowledge Panel" on the search result page, which summarizes the information on a single business in a prominent information box next to the organic and sponsored search results. Alternatively, Google may display so-called "Local Packs", which appear on top of the search result page. Local Packs display a limited number of businesses together with their contact information such as the address and telephone number, opening hours, customers' ratings and reviews, and store location on a map. Due to their prominent position on the search result page, inclusion in Knowledge Panels and Local Packs can drive significant virtual and physical traffic to local businesses.[54] The criteria and approaches for obtaining a high-ranking slot within these formats have therefore become popular issues in search engine optimisation guides.[55]

One often suggested approach to achieve a high-ranking position in local search is to provide Google with additional information about the local business.[56] While Google also collects information from public registries and firms' websites, proprietary data sources such as the Google My Business listings are found to be among the most important ranking factors for local search.[57] Businesses are compelled by Google to register a My Business account with their platform and enter complete and accurate information, such as the address and opening hours. Next to textual descriptions, businesses are encouraged to create content that encourages interaction of consumers, such as photos, videos and additional media content. This data is directly uploaded to and stored at servers of the search engine. Moreover, firms may engage with customers by responding to their reviews and questions at the search engine website to improve measures of customer satisfaction, which in turn lead to higher ranking positions. This allows Google to collect additional data on a firm's reputation and its interaction with consumers directly from its platform.

Recently, Google has integrated additional services such as "Reserve with Google" into local search.[58] When businesses sign-up for these services, consumers can directly book an appointment or make a reservation through the Google search interface. Although Google relies on third-party booking

---

[50] https://www.statista.com/statistics/297137/mobile-share-of-us-organic-search-engine-visits/, https://www.thinkwithgoogle.com/consumer-insights/mobile-search-trends-consumers-to-stores/
[51] https://www.brightlocal.com/research/local-consumer-review-survey/
[52] https://www.brightlocal.com/research/google-analytics-for-local-businesses-study/
[53] https://www.thinkwithgoogle.com/advertising-channels/search/how-advertisers-can-extend-their-relevance-with-search-infographic/ https://www.thinkwithgoogle.com/consumer-insights/mobile-search-trends-consumers-to-stores/
[54] https://moz.com/blog/ultimate-cheat-sheet-google-knowledge-panels
[55] See, for example, https://moz.com/blog/2020-local-seo-success
[56] https://support.google.com/business/answer/2721884?hl=en
[57] https://moz.com/local-search-ranking-factors, https://support.google.com/business/answer/7091?hl=en
[58] http://blumenthals.com/blog/2018/10/22/reserve-with-google-makes-its-way-into-the-3-pack-serp-on-google/

tools, the transaction process is kept within its platform. Thus, Google may access information that is exchanged between consumers and businesses on an individual and fine-granular basis.[59] Next to information on the inventory of a business and the individual booking information itself, this may also comprise data on payments, which can be executed directly over the Reserve with Google service.[60]

Besides, to organic listings in local search, Google has introduced local services ads that offer businesses a prominent ranking position in return for a monetary payment. Interactions and transactions between consumers and business that originate from these ads are monitored closely by the search platform. Specifically, phone calls and messages from users that want to engage with a local business must go through Google as the intermediary. Google states that this data may be used for spam and fraud detection, but also for "improv[ing] quality and support policies and research".[61] Transactions that are carried out over the Google platform can be reviewed by consumers and the originating ratings are displayed as so-called verified reviews, which then may contribute to a better ranking position in organic local search.

Third-party business data: Ranking criteria can incentivise third-parties to provide more data to a search engine. This can improve overall search quality. The direct access to users and the knowledge about their current information need can make the integration of third-party processes commercially attractive. In turn, search engine operators may collect data on interactions and transactions between users and businesses on their own platform.

2.1.3.2 Location-based and taste-based personalisation

For the personalisation of local search results, search engines also use additional data sources. Specifically, Google considers stated preferences that users have shared with the company such as dietary restrictions or favourite cuisines as well as revealed preferences from users' past ratings and interactions with businesses.[62] To this end, Google has access to various additional user data from its other services that are relevant for local search, e.g., geographic tracking data from Google Maps. At the same time, search engine optimization specialists conjecture that quality signals from data on behavioural engagement of consumers matter even more for a local search than for web search.[63] For example, tracking users' requests for driving directions calls to businesses and credit card transactions may present opportunities to create rich feedback data, which could then be fed back into the search algorithm as a ranking signal.

Based on the GPS location records of Android smartphone users, Google has access to a particularly extensive and fine-granular data set of individuals' geographic movement data.[64] Location records may be collected by Google even if the GPS functionality is turned off by users.[65] Even if such location information is not connected to individual user profiles, pseudonymized or anonymised data can still be valuable if data sets are sufficiently large to learn average or group-specific movement patterns and preferences.

Geographic tracking data: Tracking data from mobile devices and services allow for the analysis of user behaviour outside in physical spaces. This data can be used to learn individual level or population-level preferences.

2.1.4 Search advertising

Online search engines are mostly advertising-financed. Because search queries reveal users' information needs, advertisements can be targeted to users' interests by matching sponsored search results to query keywords. Slots for sponsored search results are usually allocated to advertisers by running auctions for each keyword, where higher bids are generally rewarded with higher-ranking slots. However, search engines regularly adjust the ranking of bids based on quality measures to

[59] https://developers.google.com/maps-booking/guides/end-to-end-integration/overview
[60] https://developers.google.com/maps-booking/guides/payments/enabling-payments
[61] https://support.google.com/google-ads/answer/7496727
[62] https://support.google.com/maps/answer/7677966
[63] https://moz.com/blog/2018-local-search-ranking-factors-survey
[64] https://apnews.com/828aefab64d4411bac257a07c1af0ecb
[65] Schmidt, D. C. (2018). Google Data Collection. Available at https://digitalcontentnext.org/wp-content/uploads/2018/08/DCN-Google-Data-Collection-Paper.pdf

ensure that users see relevant results. These quality measures usually include the expected click-through rate based on historic measurements and estimates of the ad relevance to specific keywords to maximize users' engagement with the displayed advertisements and, hence, to maximize the search engine's advertising revenues.

2.1.4.1 Improving the matching of advertisements

Collected data on user interaction can be used to improve the efficiency of the matching between advertisers and keywords in various ways. For example, while advertisers can manually choose the keywords they want to bid on, search engines offer automated matching functions, which find similar or related keywords, so that advertisements can be included in an extended set of keyword auctions. By this query expansion technique, search engines can ensure that more search queries are matched with sufficient ad keywords, which increases ad revenue. For example, Ordentlich et al. (2016) describe a "broad matching" algorithm that was implemented at Yahoo's search advertising platform and which matches ad keywords to queries based on semantic similarity learned from search query logs and users' clicks. In this machine learning approach, a neural net learns vector representations for queries, search results and advertisements based on training data from search query logs. These vector representations can then be exploited to find related queries for an ad keyword that might otherwise not be found by conventional similarity metrics.

Grbovic et al. (2015) and Ordentlich et al. (2016) find that this broad matching approach significantly outperforms conventional mechanisms with respect to the relevance of the related queries suggested by the algorithm and also with regard to the ad coverage. This performance gains directly translate into larger advertising revenues for the search platform due to higher click-through rates and improved ad allocation to more search queries. The performance improvement can be especially attributed to the neural net's capability to learn hidden relationships between keywords and the clicks of users from the historic data on user behaviour. Therefore, the algorithm can identify matching relationships that do not rely on a similar meaning or a similar syntax of keywords, but also take into account semantic similarity based on user behaviour.

To this end, larger data sets of search queries with data on user interaction are found to further improve results. Grbovic et al. (2015) learn vector representations for more than 45 Million search queries based on a training data set that comprises over 12 Billion search sessions from Yahoo Search in the US. However, Ordentlich et al. (2016) conjecture that scaling the number of vector representations even far beyond 200 Million is likely to be associated with significant performance gains as this increases the coverage of search queries for which vectors can be trained. In a field tests with live web search traffic, they find that increasing the number of vector representations from 50 Million to 133 Million increases revenue per search by almost 10% (Ordentlich et al., 2016).

2.1.4.2 Access to data to evaluate advertising effectiveness

Advertisers must monitor and analyse the effectiveness of their advertisements to manage their search advertising campaigns and thus require access to information about the allocation of ads, keyword bidding and user engagement. Thus, most search engines offer advertisers access to web portals, which, besides campaign management support, also provide functions for advertising performance evaluation. Alternatively, advertisers may access information also through application programming interfaces. By determining the scope and granularity of evaluation data that advertisers can access, search engines determine the extent of information that can be extracted. Regularly, search engines do not provide advertisers with access to data on an individual user basis, for which they cite privacy reasons as justifications.[66] In consequence, search engines have usually access to larger, more fine-grained and personally identifiable (or at least pseudonymized) data than the respective advertiser.

In general, the search engine may infer several insights about the business performance of an individual advertiser and the respective market based on the information that advertisers submit for the bidding and placement of search advertisements. For example, information such as dynamic demand patterns or profitability margins may be inferred on a fine-granular basis from the selected keywords and the submitted bids. For specific advertising techniques such as for retargeting of advertisements, which allows showing sponsored search results for products that users may have looked at in the past, the advertiser must share additional information with the search engine

[66] https://safety.google/privacy/ads-and-data/

operator. For example, for the search engine to recognize what a user has seen before, advertisers must integrate code tags of the search provider into their website. In consequence, the respective operator may be able to collect additional data on consumer behaviour beyond the scope of its search engine and other own services.

Data for search advertising: Data on search queries and user behaviour can be used to improve the matching of search advertising and to evaluate advertising effectiveness. Advertising data and integration of additional third-party data can be used by the search platform to infer business information.

2.1.5 Summary

Online search engines rely on the collection and processing of data resources to retrieve relevant information from the distributed content and documents of the World Wide Web. Moreover, data plays a key role in improving the quality of search results, which is determined by the ranking decisions of a search engine.

Crawling of web content and creating the *search index data* represents the basis for the matching of users' search queries to relevant websites. Web index data can be collected from publicly accessible information on websites. Completeness and freshness of the web index data determine the set of search results that can be retrieved by a search engine.

Search query data provides the basis for improving search quality by understanding users' search intent and providing them with suitable results. Combined with *data on user behaviour,* search query logs are used to analyse the implicit quality feedback given by users' observed decisions and actions on the search engine platform. Collection and analysis of this data are used to improve the ranking decisions and the matching of search queries. Augmented search query logs may exhibit different levels of *data granularity*. For example, the data may contain only search queries or search queries and the final search result clicked by a user, or search queries, the user clicks and associated search result pages. A finer granularity will generally allow to infer more information and thus lead to larger improvements in search quality. Furthermore, these data sets may be entirely *anonymous*, i.e., it is then unknown which log entries refer to the same users, or *pseudonymized*, which allows different search request of the same user to be identified and connected.

Personally identifiable *individual user data* can be used to infer the context of a search request and thus to improve search results by personalising the ranking of search results. Collecting data on users' behaviour can reveal information that a search engine may not obtain through explicit feedback from a user (e.g., the search query). User profiles may be created from observing user behaviour at the search engine website, but also from tracking user activity on other services and in other domains. Inferred information from this data can improve the quality of search results, especially concerning new search queries. Ranking information may also be retrieved by analysing the similarity of user profiles and their past behaviour. Collection of *geographic tracking data* extends the depth of user profiles by including information on user behaviour in physical environments. This can especially improve the quality of local search.

Ranking criteria based on the accuracy and completeness of third-party-provided information may create incentives for businesses to create *third-party business data* directly at the search platform. Such incentives are particularly strong for a search engine with a large user base that attracts third-parties through indirect network effects. On the one hand, such data provided by third-parties can improve the quality of search results, as provided information is verified, processed and possibly customised by businesses. On the other hand, directly provided data reduces the costs of data collection for the search engine, which allows for the collection of accurate data on a larger scale. Finally, data creation may establish switching costs for third-parties if the entered data is not portable or compatible. Moreover, search engines as intermediaries may collect data on interactions and transactions carried out between users and businesses on their platform. This data can be used for improving search quality, but also for use in complementary markets.

Finally, search query data and data on user behaviour may be used to improve the matching quality and the coverage of search advertisements to search query keywords. Tracking advertising effectiveness on third-party websites can give search engine access to additional data on user

behaviour. Moreover, the search engine can determine the level of access to data that advertisers use for performance evaluation of their search advertisements.

2.2 E-commerce

Concerning the role of data in e-commerce, we focus on (i) demand forecasting, which is crucial for the curation of product portfolios as well as for efficient operations and logistics, and (ii) recommendation systems, which personalise a consumer's shopping experience and represent a now widespread mechanism to facilitate consumers' discovery of new and suitable products in virtual e-commerce markets. Next to these two major applications data is also used for dynamic pricing, which enables retailers to automate price setting taking into account competitors' prices, products' profit margins, inventory and other factors.[67] Moreover, personally identifiable data on transactions can be used to create or enrich deep user profiles that can be leveraged for targeted marketing activities and personalised pricing. In this context, media services and interactive assistant services that may be offered on top of physical products can collect additional data on user behaviour and consumer preferences.

2.2.1 Demand forecasting

2.2.1.1 Product portfolio curation and efficiency of operations

The curation of a firm's product portfolio offered to customers is a major strategic variable and a key differentiator in the retailing business. Developing their product portfolio, retailers aim to offer items that match consumers' demand and yield high-profit margins, but they must also consider complexity and high costs for product fulfilment from product variety (Jiao and Zhang, 2005[68]). With the advent of electronic marketplaces, online retailers can access much more data on consumers' usage and purchases, which allows these firms to base product portfolio decisions on data-driven demand forecasts and consumer preferences inferred from behavioural user data. Rather than relying solely on traditional marketing research methods such as conjoint-based analyses of consumers' stated preferences, online retailers can exploit aggregate data about consumers' revealed preferences at their website to make decisions on introducing new products or abandoning existing offers. For example, firms may analyse observed data on consumers' decisions to view or buy specific products to infer popular product attributes across items (see, e.g., Tucker & Kim, 2009[69]) or detect trending interests among customers (e.g., in the case of fashion items). In the context of digital goods, access to individual-level usage data from intelligent, connected devices can offer additional insights on consumers' interaction and satisfaction with items after the products have been purchased (Porter & Heppelmann, 2015[70]). For example, Amazon uses data on consumers' reading behaviour from Kindle devices to identify popular and exciting books (e.g., the company promotes "books Kindle readers finish in three days or less").[71] Aggregate usage data can thus be used to better tailor the offered product mix to consumers' interests, but also to inform the design of entirely new products (see, e.g., Hou & Jiao, 2020[72]).

Next to the offered product portfolio, immediate availability and timely delivery of products to end consumers are key competitive factors in e-commerce. Besides, efficient warehousing and logistics can save inventory and transportation costs. In this context, accurate demand prediction is the basis for efficient supply chain management including automated ordering, in-stock management and facilities planning (Seeger et al., 2016[73]). Specifically, "probabilistic demand forecasts are crucial for having the right inventory available at the right time and in the right place" (Salinas et al., 2019[74], p.1). In contrast, if forecasts underestimate actual demand, retailers will miss out on short run sales

---

[67] https://www.businessinsider.de/international/amazon-price-changes-2018-8/?r=US&IR=T, https://dzone.com/articles/big-data-analytics-delivering-business-value-at-am

[68] Jiao, J., & Zhang, Y. (2005). Product portfolio planning with customer-engineering interaction. *IIE Transactions*, *37*(9), 801-814.

[69] Tucker, C. S., & Kim, H. M. (2009). Data-driven decision tree classification for product portfolio design optimization. *Journal of Computing and Information Science in Engineering*, *9*(4), 1-14.

[70] Porter, M. E., & Heppelmann, J. E. (2015). How smart, connected products are transforming companies. *Harvard Business Review*, *93*(10), 96-114.

[71] https://www.theguardian.com/technology/2017/may/26/amazon-new-york-bookstore

[72] Hou, L., & Jiao, R. J. (2020). Data-informed inverse design by product usage information: a review, framework and outlook. *Journal of Intelligent Manufacturing*, *31*(3), 529-552.

[73] Seeger, M. W., Salinas, D., & Flunkert, V. (2016). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* (pp. 4646-4654).

[74] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2019). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*.

and profits from going out of stock (Bajari et al., 2018[75]). On the other hand, if forecasts exceed actual demand, retailers may need to markdown or liquidate overstock products.

To improve the accuracy of their forecasts, electronic marketplaces leverage data on consumer demand that can be collected from users' current behaviour and past purchases. Collecting historic demand data can be informative for future sales, as seasonal trends or correlations with external events or other products may systematically influence demand patterns. Demand and supply chain data is often highly intermittent (i.e., there are zero sales in some periods) and bursty (i.e., there is high turnover in a few periods), which makes forecasting in e-commerce more challenging than in other domains, but also renders accurate predictions economically more valuable.
Longer timeseries data on an individual product, i.e. observing demand for a product over a longer time horizon, may thus contribute to better medium- and long-term predictions of future sales. Moreover, timeseries data from related products may be used to improve prediction accuracy, especially when retailers sell a large number of such related products (Salinas et al., 2019). Learning from related items can be especially valuable for forecasts of new items with a short demand history (Seeger et al., 2016, Salinas et al., 2019). Amazon researchers have recently proposed a machine-learning approach that learns a global prediction model from historical timeseries data of all products in a dataset and can thereby significantly improve the forecast accuracy of product sales as demonstrated based on a large product dataset from Amazon (Salinas et al., 2019). Similar performance gains have been found for the forecasting of sales for the online marketplace of Walmart based on a prediction model that takes into account sales correlations and relationships between products according to the product hierarchy (Bandara et al., 2019[76]). Next to information about historic purchases, observed data on consumer behaviour such as the search volume or page views of a product can be used to improve demand forecasts (see, e.g., Yan et al., 2014[77]). Moreover, online retailers may also analyse user-generated data such as product reviews to predict future demand of the respective product. For instance, it has been found that review volume and ratings can serve as predictors of future sales (Chong et al., 2017[78]; Li et al., 2016[79]).

Beyond aggregated sales predictions, online retailers aim to forecast sales on a more fine-granular basis, e.g., for groups of customers of a specific demographic or within a specific geographic region, and even at the individual level. In 2013, Amazon obtained a patent on "anticipatory package shipping" designed to deliver products to local areas even before customers would purchase the product.[80] Forecasting of individual needs requires the collection of deep data to accurately detect repeating purchasing patterns of customers and to infer consumers' current and evolving interests. In this context, data collection through home automation devices and voice assistants can expand retailers' ability to learn information about consumers' context and habits beyond users' behaviour on virtual shopping sites.

Demand forecasting data: Collection and analysis of aggregate and individual-level behavioural user data (product interactions and purchases) can improve demand forecasting, which is used for product portfolio decisions and improving the efficiency of operations.

Operators of electronic marketplaces as intermediaries have access to data on the transactions of third-party businesses that are executed over their platform.[81] By providing ancillary services such as fulfilment (e.g., storage, logistics, and delivery), marketplace operators can collect further data

[75] Bajari, P., Chernozhukov, V., Hortacsu, A., & Suzuki, J. (2018). The Impact of Big Data on Firm Performance: An Empirical Investigation. Working Paper.
[76] Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing* (pp. 462-474). Springer, Cham.
[77] Yuan, H., Xu, W., & Wang, M. (2014). Can online user behavior improve the performance of sales prediction in E-commerce?. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2347-2352).
[78] Chong, A. Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, *55*(17), 5142-5156.
[79] Li, B., Ch'ng, E., Chong, A. Y. L., & Bao, H. (2016). Predicting online e-marketplace sales performances: A big data approach. *Computers & Industrial Engineering*, *101*, 565-571.
[80] US Patent 8,615,473 B2, see, e.g., https://techcrunch.com/2014/01/18/amazon-pre-ships/.
[81] There is an ongoing formal investigation by the European Commission into Amazon's dual role as an upstream marketplace intermediary and a downstream retailer that is scrutinising Amazon's use of data with regard to third-party businesses. See, e.g., http://competitionlawblog.kluwercompetitionlaw.com/2018/11/30/the-eus-competition-investigation-into-amazon-marketplace/?doing_wp_cron=1588254455.2190229892730712890625

on the operations of third-party businesses, e.g., on the sourcing of products. Offering these services not only for sales on their marketplace allows intermediaries to also gather information on transactions carried out outside of their platform. Likewise, consumer-directed offerings such as payment and billing services can be leveraged to track transactions of consumers outside of the marketplace platform.

Data on third-party businesses: Marketplace operators may observe data from transactions of third-party sellers, which may contain commercially valuable information. Ancillary platform services for third-parties can extend the scope of data collection on third-parties.

2.2.2 Personalised recommendations

Today, large electronic marketplaces offer consumers huge product catalogues with up to hundreds of millions of products.[82] While digital technology has dramatically reduced transaction costs for searching and comparing products in online markets, consumers' ability and time to process the available product information is limited. Thus, in e-commerce, consumers are often assisted by automated product recommendation agents to discover the products that fit their needs and preferences. State-of-the-art recommender systems exploit consumer, product and transaction data to provide consumers with personalised recommendations that are tailored to a buyer's context.

Online retailers and marketplaces like Amazon – which was among the first e-commerce businesses to implement a recommendation system on a large scale – aim to offer consumers an individually personalised shopping experience based on consumers' inferred interests (Smith & Linden, 2017)[83]. To this end, consumers are presented with personalised product recommendations in various formats in different places on a retailer's website during each step of the purchasing process. For instance, consumers are offered products based on "Related to Items You Have Viewed", "Inspired by Your Browsing History" and "Customers Who Bought This Item Also Bought".[84] According to McKinsey (2013) 35% of sales on Amazon come from product recommendations.[85] Already in 2003, Amazon engineers stated that click-through rates and conversion rates of recommended products "vastly exceed those of untargeted content" (Linden et al., 2003, p.76)[86].

In order to provide consumers with relevant and accurate suggestions, recommendation systems analyse the interdependencies between products that have been viewed, liked or purchased in combination by other consumers (Smith & Linden, 2017). To this end, recommendation systems process and analyse large sets of product and consumer data including product descriptions and product attributes, consumers' purchases and shopping behaviour, and product ratings from consumers. Based on this data, recommendation systems then predict consumer-specific virtual ratings for other available products in order to decide which items suit a specific consumer best. From an economic perspective, recommendation systems are effective if they increase the number of sales by facilitating the discovery of new products, raise the cross- and up-selling of additional products or enhance consumer satisfaction by improving the shopping experience. To achieve these economic goals, it is necessary that consumers perceive the recommendations as relevant and to be of high quality. To measure quality from a technical perspective, recommendation systems are mainly evaluated based on their accuracy, i.e., how closely recommendations and predicted ratings of products match the real preferences of users.

2.2.2.1 Recommendation Algorithms

Many popular recommendation systems in the e-commerce sector implement *collaborative filtering* algorithms, which identify new user-product relationships based on the analysis of relationships among users and interdependencies between products (Koren et al., 2009)[87]. Specifically, most

[82] https://www.scrapehero.com/number-of-products-on-amazon-april-2019/, https://www.businesswire.com/news/home/20160614006063/en/Products-Amazon-Carry-Categories
[83] Smith, B., & Linden, G. (2017). Two decades of recommender systems at amazon. com. *IEEE Internet Computing*, *21*(3), 12-18.
[84] https://dzone.com/articles/big-data-analytics-delivering-business-value-at-am
[85] https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers
[86] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76-80.
[87] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30-37.

commercial systems are based on *item-based nearest-neighbour models* (Koren & Bell, 2015)[88], which recommend products that are most similar to products that a consumer has liked in the past. However, product similarity according to this notion is not derived directly from product attributes, but from the transactions and interactions of other consumers with these products. Alternatively, *user-based models* directly exploit the similarity of consumer profiles and their transaction histories to recommend products that other users have liked in the past. Finally, *latent factor models* leverage the similarity along both the user and the product dimension to identify new user-product relationships. By using matrix factorization methods, these models characterize product and consumers according to an aggregated set of latent factors, which are computationally inferred from existing user-product patterns in the data (Koren et al., 2009). Due to the high expressive ability of latent factor models to describe data and capture correlation patterns irrespective of the application domain, they are found to often provide the most accurate results (Koren & Bell, 2015, p. 94; Aggarwal, 2016[89]).

Collaborative filtering approaches are conceptually based on the processing of a *user-product rating data matrix,* which is illustrated conceptually in Figure 1. Hence, these algorithms require an input data set that contains the set of users (the number of columns in the data matrix depicted in Figure 1), the set of products (the number of rows in the data matrix) and the interdependencies between users and products, expressed as numerical or binary ratings of users for products (the cell entries in the data matrix). Because most consumers have only bought or interacted with a small share of the overall available products, rating entries in the matrix are *sparse*. That is, most cells of the user-product matrix are empty because no data on the respective user-product relationship is available.[90] For example, in Figure 1, product A has only received one user rating and for users U_1 and U_4 only a single product rating is available, respectively. Hence, it is difficult to identify similar products to A and to derive accurate recommendations for user U_4. In consequence, the accuracy of recommendations hinges critically on a firm's ability to reduce sparsity by collecting data and to implement algorithms that maximize the information that can be deduced from sparse data sets.

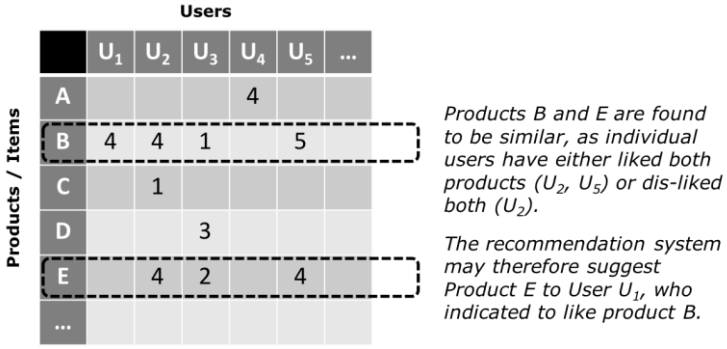

Figure 1: Illustration of a user-product rating data matrix. Cell entries represent users' ratings of the respective product on a rating scale between zero and five.

A major challenge for recommendation systems specifically arises when accurate recommendations must be retrieved for new consumers that just registered with an online retailer or for new items that were just added to the product catalogue. Because new users have not yet purchased any products and may also not have interacted with the online shop, rating entries for such a user are all empty. Likewise, a new product has not yet received any ratings from users. In consequence, a collaborative filtering algorithm is unable to find similar products or users on whose basis recommendations could be generated. This gives rise to the well-known *cold-start problem* of recommendation engines, which can arise for new users as well as new products. A naïve approach to alleviating the cold-start problem is to randomly include new products in the set of recommendations and to thereby explore new user-product relationships and generate ratings in the form of user feedback. This may lead to low customer satisfaction if the recommended product was

[88] Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In *Recommender Systems Handbook* (pp. 77-118). Springer, Boston, MA.
[89] Aggarwal, C. C. (2016). *Recommender Systems*. Cham: Springer International Publishing.
[90] To illustrate the degree of sparsity in real-world data sets, one may consider the user-item data matrix made available by Netflix during the recommendation challenge in 2006, which consisted of a user-movie data set where only every 85th cell contained a rating in a matrix with 8.5bn entries (Funk, 2006).

perceived to be a bad match or of low quality. Instead, concerning new users, data from other sources such as demographic information may be used to infer interests of consumers and to provide initial product recommendations. In addition, deeper user data make it more likely that similar users can be identified based on larger user profiles and thus can mitigate the cold-start problem for new customers.

Consumer-product data: Personalised recommendation systems require data on the user base and the product catalogue of an e-commerce retailer. Recommendation accuracy increases with the number of consumers' purchases and product interactions that the system can collect data on. A minimum number of data points is necessary to overcome the cold-start problem of recommendation systems.

Concerning new products, *content-based recommendation algorithms* are a widely applied approach to alleviate the cold-start problem of collaborative filtering methods. Content-based algorithms identify similar products based on the descriptions and properties of the products themselves and may also take into account user-generated tags or textual reviews. Products are then recommended to users who have high ratings for very similar products. While recommendations of content-based systems are robust against popularity bias (i.e., recommending only mainstream products), they allow for less discovery of really new and even surprising products that a consumer might have not thought of herself (*serendipity*). Moreover, content-based recommendations are less *diverse* than in the case of collaborative filtering methods, because the similarity metric is directly tied to the product attributes. Therefore, in practice, most recommendation engines implement hybrid systems that take into account the signals of both content and collaborative filtering algorithms to benefit from the relative strengths of both approaches.

2.2.2.2 Data collection: explicit and implicit user feedback of products

Collaborative filtering systems require data on consumers' ratings and preferences for products, which reduces sparsity of the item-user matrix and improves recommendation accuracy. In academic studies of recommendation algorithms, consumers' ratings of items are often assumed to be given as explicit user feedback. For example, customers in online shops are regularly asked to rate a product on a numerical scale after purchase. Similarly, customers may give binary feedback by indicating whether or not they like a specific product. A well-known characteristic of explicit ratings is their long-tail distribution in commercial systems (Aggarwal, 2016, Cremonesi et al. 2010[91]). For example, in the internal rating data set that was published by Netflix in 2006, the 1.7% most popular movies account for 33% of all ratings (Cremonesi et al. 2010). However, recommending already popular items provides relatively low value to users. Thus, to allow for genuine product discovery, rating data on less popular products is much more valuable for recommendations.

Data on user ratings (explicit feedback): Online retailers may collect ratings for products directly from their consumers. This explicit feedback data reduces the sparsity of the consumer-product data matrix and improves recommendation accuracy.

In practice, user ratings are often inferred from indirect feedback, i.e., the behaviour of consumers (see, e.g., Ekstrand et al., 2011[92]; Hu et al., 2008[93]). Observed user behaviour such as the number of website visits, the time spent on websites, clicking behaviour and interaction with a website's content can reveal users' valuation and the relative preference ranking for products (Koren et al., 2009). In this context, studies have demonstrated that already the decision of users to view, purchase or rate a specific product conveys valuable information with respect to a user's preferences (Aggarwal, 2016). While this type of observed or inferred data is likely to be noisier than explicit user feedback, it may contain additional information that users themselves cannot or are not willing to disclose explicitly (Ekstrand et al., 2011). Besides, implicit ratings are less susceptible to malicious attacks such as fake ratings. Implicit feedback data may not only be used to generate rating

[91] Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 39-46).
[92] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends in Human–Computer Interaction*, *4*(2), 81-173.
[93] Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 263-272). IEEE.

estimates but also be processed such that they are associated with confidence levels on the reliability of the generated estimates (Hu et al., 2008).

Data on user behaviour (implicit feedback): Instead of or in addition to explicit feedback, online retailers can collect implicit feedback data from observed user behaviour. Implicit feedback allows inferring consumers' preferences and ratings for products based on their revealed decisions and actions. Like explicit feedback, implicit feedback data reduces the sparsity of the consumer-product data matrix and improves recommendation accuracy.

Online retailers frequently make use of both explicit and implicit user feedback (Ekstrand et al., 2011). Especially implicit feedback is "self-generating" as a by-product of consumers' usage of the e-commerce service through either the retailer's website or application. Retailers with a larger customer base will, therefore, have access to more data from which ratings can be inferred, everything else being equal. For example, eBay's recommendation system has been upgraded in several countries to take into account a similarity metric for products that are based on information extracted from users' historic search queries (Katukuri et al., 2014[94])

Content-based recommendations require no data on users nor information about the relationships between users and products. However, recommendations that exclusively rely on the similarity of products are often less accurate and are often perceived as rather boring by users, because genuinely new, although still accurate recommendations that match a users' interest are missing. That is due to the lower diversity and serendipity of content-based systems. Moreover, collaborative filtering methods scale more efficiently with an increasing number of available products. Both reasons have been cited for Amazon's early decision to implement collaborative filtering systems (Linden et al., 2003). Thus, data on content similarity may be important to overcome shortcomings of collaborative filtering, especially the cold-start problem, but without access to data on the interdependencies between users and products, the effectiveness of recommendation engines will be significantly worse.

Data on product characteristics: Data on product characteristics can be used to identify similar products for recommendations and can thus alleviate the cold-start problem of recommendation systems.

2.2.2.3 Data collection: user profiles, dynamic user behaviour and cross-domain data

Next to implicit and explicit feedback data, firms that provide recommendations frequently collect additional data on users and products. This is mainly to improve recommendation quality by reducing sparsity and to address the cold-start problem for new consumers or new products or to establish a data basis for alternative algorithmic approaches if collaborative filtering is not directly applicable, especially if insufficient consumer-product observations are available. For example, in the case of the online marketplace eBay, the interaction of consumers with product items is short-lived, because most offers are only valid for a limited time (Chen and Canny, 2011[95]).

Especially for new users for which no rating or implicit feedback data is yet available, the common strategy is to collect and incorporate additional individual-level information about users themselves. For example, user attributes such as demographics can be used to infer recommendations based on what users with similar attributes have liked (Koren et al., 2009). More generally, user data is found to be effective in adjusting for systematic biases of user attributes on ratings and thus the incorporation of user data regularly improves recommendation accuracy.

While basic recommendation algorithms rely on static rating data for consumer-product relationships, scholars and practitioners have long acknowledged the value of temporal data. The collection of information on when implicit or explicit feedback is obtained from a consumer can achieve large quality improvements. Firstly, because consumers' preferences for products are not static as user taste evolves or new products are introduced, more recent data points contribute more value to accurate recommendations (Koren, 2009). Therefore, continuous data collection that obtains fresh data on users' preferences constitutes an important input to recommendation algorithms. Secondly,

[94] Katukuri, J., Könik, T., Mukherjee, R., & Kolay, S. (2014, October). Recommending similar items in large-scale online marketplaces. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 868-876). IEEE.
[95] Chen, Y., & Canny, J. F. (2011). Recommending ephemeral items at web scale. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 1013-1022).

the recognition of repeating patterns in user behaviour can be leveraged to improve recommendations. Thirdly, the timestamp of an explicit rating relative to the time of the actual product purchase can already convey implicit information on the customer's satisfaction with the product.

Recently, deep learning techniques have been proposed to also integrate additional behavioural data sources to improve the performance of recommendation systems. For example, recurrent neural networks can be integrated into recommendation systems to leverage browsing data and the sequence of page views within a consumer's browsing session to enhance short-term recommendations (Batmaz et al., 2019). Convolutional neural networks can be used to extract latent factors, which can be used for the computation of similarity relationships, from text data or even other media such as images or audio when these factors cannot be obtained directly from the feedback of users. Zhou et al. (2016)[96] demonstrate that recommendations can integrate visual interest profiles of users, i.e., such recommendation systems can account for the role of product images and their perception by individual users when inferring consumers' preferences for new products.

Fine-granular data on user behaviour: The accuracy of personalised recommendations can be improved by the use of temporal data that capture dynamic user behaviour and by inferring information from fine-granular observed data on user behaviour.

An additional approach to improve recommendation quality is the collection of user data from other domains. External data sources are especially suited to overcome the user cold-start problem by expanding access to more personal data of users. Specifically, studies have demonstrated that user data from social networks can be incorporated to improve recommendations. Shapira et al. (2013)[97] demonstrate that data from Facebook such as a user's 'likes' can partially substitute rating data and increase the accuracy of recommendations. Besides, experimental results indicate that data from Facebook user profiles can improve users' satisfaction with product recommendations (Heimbach et al., 2015[98]). Moreover, social networks allow retailers to retrieve information on users' social relationships and contact networks. This information can be used to predict the similarity of users based on their social ties (see, e.g., Ma et al., 2011[99]). By analysing data on social relationships, recommendations can be further differentiated according to different levels of trust between users. In practice, for example, media services such as Netflix use social data, especially friend connections to personalise their systems (Amatriain & Basilico, 2015).

Individual user data: Information on individual users from other domains and services can be used to overcome the user cold-start problem and improve recommendation accuracy.

Beyond sourcing of individual user information from other domains, cross-domain recommendation systems exploit patterns of similar user interests or behaviours in other areas to expand the data basis on user-item relationships and to infer similar user profiles based on a wider set of activities (Cantador et al., 2015[100]). For example, Elkahky et al. (2015)[101] collect behavioural user data from four different domains (web search, news consumption, mobile app usage, movies and TV) to train a deep learning model and demonstrate that additional data from other source domains improves prediction accuracy in a specific target domain. The gains in recommendation quality are particularly pronounced for new users that have very few or no data points in a single domain. In the e-commerce context, cross-domain recommendations are especially relevant as large online retailers offer a wide

[96] Zhou, J., Albatal, R., & Gurrin, C. (2016). Applying visual user interest profiles for recommendation and personalisation. In *International Conference on Multimedia Modeling* (pp. 361-366). Cham: Springer.
[97] Shapira, B., Rokach, L., & Freilikhman, S. (2013). Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction*, *23*(2-3), 211-247.
[98] Heimbach, I., Gottschlich, J., & Hinz, O. (2015). The value of user's Facebook profile data for product recommendation generation. *Electronic Markets*, *25*(2), 125-138.
[99] Ma, H., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011). Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 287-296).

[100] Cantador, I., Fernández-Tobías, I., Berkovsky, S., & Cremonesi, P. (2015). Cross-domain recommender systems. In *Recommender Systems Handbook* (pp. 919-959). Springer, Boston, MA.
[101] Elkahky, A. M., Song, Y., & He, X. (2015). A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 278-288).

variety of products and services. Thus, consumers may buy physical goods from a retailer, but also consumer media services, such as video or audio streaming (Cantador et al., 2015).

The expansion of e-commerce platforms into adjacent markets and the bundling of products and services now allow some retailers to collect a large variety of user data from different domains. For example, Amazon's smart home device Echo and its voice assistant Alexa are designed to accompany consumers in their everyday life and may, therefore, be able to collect data and track user behaviour not only in online environments but also in physical spaces. Amazon processes billions of interactions a week based on their Alexa system, which generates data on consumers' schedules, locations and preferences.[102] Google already uses data from its voice assistant to personalise advertisings at websites or smartphone apps.[103] Moreover, streaming services such as Amazon Prime Video and Amazon Music allow the company to create user profiles for customers based on their preferences for different types of media content and infer similarity of consumers based on a wide product and service catalogue. Other retailers like Walmart have combined proprietary data such as customer purchasing data with publicly accessible web data of consumers. The so-called Social Genome project is aimed at inferring consumer tastes and similarities from other social network services and domains than e-commerce.[104]

Data on cross-domain user behaviour: Data on user behaviour in other domains and on other services can be used to infer more general preferences and new similarity relationship between users. This can create additional user-product relationships and thus improve recommendation accuracy, especially for new users.

2.2.2.4 Data for machine learning algorithms and additional personalisation techniques

More recently, deep learning algorithms have been suggested to further improve the accuracy of recommendation systems and to address scalability and cold-start issues (Batmaz et al., 2019)[105]. Better recommendation performance of these techniques is achieved by reducing sparsity of the user-product matrix (see, e.g., Sedhain et al., 2015[106] and Wang et al., 2015[107]) and by integrating additional data sources, such as fine-grained browsing data of customers and information on the evolution of user tastes, to train and calibrate machine learning models (see, e.g. Wu et al., 2016[108]). By doing so, these methods can improve the prediction of customers' future consumption and increase recommendation accuracy as well as coverage of products that are included in recommendations. On a fundamental level, deep learning methods enable a more composite architecture of recommendation systems, which allows for the extraction and integration of information from various heterogeneous data sources (Batmaz et al. 2019). Finally, these algorithms scale especially well by employing feature extraction and dimensionality reduction techniques and are thus well-suited to handle large datasets (Batmaz et al. 2019, Elkahky et al. 2015).

Machine learning techniques have also been implemented to address specific tasks that arise in the context of personalised recommendations. For example, consumers may use the search interface of an online retailer and enter a search query when looking for a product. Instead of relying on general relevance metrics and audience-wide popularity measures as in traditional information retrieval techniques (see Section 2.1), tailored search results selected from a set of personalised recommendations may be more relevant to the respective individual consumer. In addition, when displaying multiple recommended products to a user, the relative ranking of items plays an important role. Learning-to-rank algorithms take not only into account the ratings of the user-item matrix when determining the order of display, but also account for a variety of additional features such as simple item metadata, user interaction, but also signals from additional classification algorithms. Amatriain

[102] https://www.technologyreview.com/2019/11/05/65069/amazon-alexa-will-run-your-life-data-privacy/
[103] https://www.bloomberg.com/news/articles/2019-12-31/you-re-home-alone-with-alexa-are-your-secrets-safe-quicktake
[104] https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109
[105] Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, *52*(1), 1-37.
[106] Sedhain, S., Menon, A. K., Sanner, S., & Xie, L. (2015). Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web* (pp. 111-112).
[107] Wang, H., Wang, N., & Yeung, D. Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1235-1244).
[108] Wu, S., Ren, W., Yu, C., Chen, G., Zhang, D., & Zhu, J. (2016). Personal recommendation using deep recurrent neural networks in NetEase. In *2016 IEEE 32nd international conference on data engineering (ICDE)* (pp. 1218-1229). IEEE.

& Basilico, 2015 (2015, p.395)[109] report that including additional features and optimized models improved the ranking at Netflix more than 250% over simple popularity-based rankings, whereas the inclusion of ratings alone improved ranking performance only by about 40%.

2.2.2.5 Data monetization: Economic benefits from personalised recommendations

While the economic performance evaluation of recommendation systems is conceptually and practically challenging, the aggregate evidence on the business value of personalised recommendations points to significant positive effects from personalised recommendations and to continuous benefits from improved recommendation quality (Jannach and Jugovac, 2019[110]). To this end, a rich academic literature system has identified several parallel channels through which personalised recommendations create economic value. The magnitude of these benefits is often difficult to quantify, given the proprietary nature of many data sets and business performance being confidential information. However, regular contributions from practitioners of several successful firms in digital markets highlight the important role that recommendation systems play for their business models and the benefits that can be obtained from improved technical performance, especially in the context of e-commerce and media services (see, e.g., Linden et al. 2003; Chen and Canny, 2011).

[109] Amatriain, X., & Basilico, J. (2015). Recommender systems in industry: A Netflix case study. In *Recommender Systems Handbook* (pp. 385-419). Springer, Boston, MA.
[110] Jannach, D., & Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, *10*(4), 1-23.

In the academic literature, it is widely recognized that personalisation can increase *customer satisfaction* and retention (see, e.g., Ansari and Mela, 2003[111]). By reducing search costs and improving product fit, recommendations can facilitate consumers' purchase decisions (Ansari et al., 2000[112]). Because personalisation is only feasible if the service provider has access to sufficient information and data about the individual user, consumers must invest in creating this data either by manually entering volunteered data (e.g. in the form of product ratings) or by contributing observed data through the usage of the service over a sufficiently long period. In consequence, this gives rise to switching costs and higher *customer loyalty* if this user data cannot be transferred to competitors. Thirumalai & Sinha (2013)[113] find that retailers with large-scale operations that provide a greater variety of products and thus realise higher customer satisfaction by providing personalised recommendations are the most likely to benefit from higher customer loyalty due to the personalisation of customers' decision process.

Empirical studies of online retailers find that personalised recommendations generally increase *sales and revenues* at the firm and product level, although effectiveness is moderated by several additional factors (Lee and Hosanagar, 2016[114], 2018[115]). For example, De et al. (2010)[116] show that the use of a recommendation system has a significant positive effect on online sales in the case of a large retailer of women clothing. Lee and Hosanagar (2016) demonstrate that the introduction of a recommendation system increases the conversion rate by 5.9% for a US retailer. At the product level, recommendation links generally increase visibility among potential buyers and lead to more page views by consumers. However, by displaying recommendations on product pages retailers may also promote substitutes to products that shoppers are interested in, thus driving consumer demand away from the original product. Still, the net gain of these two effects is found to be significantly positive and taken together recommendations increase total sales of recommended products and its substitutes by 11% on average in the case of a US fashion retailer (Kumar and Hosanagar, 2019[117]).

Moreover, personalised recommendations are found to have a significant effect on users' consumption patterns. In this context, several empirical studies have evaluated the impact of product recommendations on a *variety of product sales*. On the one hand, recommendations may facilitate the discovery of niche products and thus increase the sales of long-tail products (Brynjolfsson et al., 2006[118], Brynjolfsson et al. 2011[119], Oestreicher-Singer and Sundararajan, 2012[120]). This typically benefits consumers as additional surplus is created when consumers can satisfy more diverse needs and product fit is improved (Hinz & Eckert, 2010[121]). In turn, this allows retailers to serve a larger user base with more diverse preferences. Moreover, as niche products are often associated with higher margins, higher sales of *long-tail products* increase a retailer's profit (Hinz & Eckert, 2010). On the other hand, recommendations may also give rise to a *superstar effect*, because blockbuster products are recommended more frequently and sales concentration is thus increased (Brynjolfsson et al., 2010[122]). Fleder and Hosanagar (2009)[123] highlight that these two economic effects can be reconciled: recommendations can facilitate product discovery at an individual level, i.e., each

[111] Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, *40*(2), 131-145.
[112] Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendation Systems. *Journal of Marketing Research*, *37*(3) 363-375.
[113] Thirumalai, S., & Sinha, K. K. (2013). To personalize or not to personalize online purchase interactions: implications of self-selection by retailers. *Information Systems Research*, *24*(3), 683-708.
[114] Lee, D., & Hosanagar, K. (2016, April). When do recommender systems work the best? The moderating effects of product attributes and consumer reviews on recommender performance. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 85-97).
[115] Lee, D., & Hosanagar, K. (2018). How Do Product Attributes Moderate the Impact of Recommender Systems?. *Available at: https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2018/10/FP0315_WP_2018Oct.pdf*
[116] De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, *56*(11), 1930-1945.
[117] Kumar, A., & Hosanagar, K. (2019). Measuring the value of recommendation links on product demand. *Information Systems Research*, *30*(3), 819-838.
[118] Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. *Sloan Management Review*, *47*(4), 67-71.
[119] Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, *57*(8), 1373-1386.
[120] Oestreicher-Singer, G., & Sundararajan, A. (2012). The visible hand? Demand effects of recommendation networks in electronic markets. *Management Science*, *58*(11), 1963-1981.
[121] Hinz, O., & Eckert, J. (2010). The impact of search and recommendation systems on sales in electronic commerce. *Business & Information Systems Engineering*, *2*(2), 67-77.
[122] Brynjolfsson, E., Hu, Y., & Smith, M. D. (2010). Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research*, *21*(4), 736-747.
[123] Fleder, D., & Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, *55*(5), 697-712.

customer on its own discovers new products through personalised recommendations, but overall diversity may still decrease because recommendations often suggest the same products. The basic rationale of this theory is confirmed by a large field experiment with a North-American online retailer by Lee and Hosanagar, 2019[124].

Recommendation systems also facilitate the *cross-selling of complementary products* (see, e.g., Pathak et al., 2010[125]) by suggesting customers additional items related to the products in their shopping baskets or to their previous purchases at different stages of the customer's shopping lifecycle. For Amazon, Oestreicher-Singer and Sundararajan (2012) find that recommendations of complementary products, which similar customers have previously purchased together, can lead to a three-fold increase of additional demand for these complementary products. Diverse co-purchase recommendations such as for items from other product categories can further increase product sales (Lin et al., 2017[126]). Carmi et al. (2017)[127] document that demand spikes for a specific product due to external events (e.g., in the case of a highly publicized book review) diffuse to products that are recommended with the original product. This can lead to substantial revenue gains for the retailer beyond the initial demand increase for the original product. Finally, recommendations can be effective in increasing sales by converting browsing users without concrete purchase intentions into buyers of recommended products (Schafer et al., 2001[128]).

With regard to consumer behaviour, experimental studies find that consumers' product choices are more likely to be influenced by personalised recommendations from recommendation systems than by human advice or by non-personalised information about other customers' purchases (Senecal & Nantel, 2004[129]). Moreover, research on behavioural effects shows that consumers' purchase decision and their willingness to pay can be influenced by the displayed ratings of recommendations (Adomavicius et al., 2018[130]). Thus, consumers seem to value recommendations as informative when forming their product preferences (Adomavicius et al., 2018), especially when the recommendation system is viewed as reliable (Adomavicius et al. 2013[131]). This is likely to be magnified in the early stages of a purchase decision when consumers face high uncertainty regarding available products and their own needs. These individual-level behavioural effects can positively affect retailers' revenues and profitability.

In practice, recommendations play a prominent role in online retailers' websites. Already in 2013, 94% of e-commerce sites considered "recommendation systems to be [a] critical competitive advantage to be implemented" (cited by Lee & Hosanagar, 2016, p.85). Especially, retailers with a large product catalogue and product variety rely on various personalisation and recommendation tools. The economic value of recommendations has also been highlighted in other domains, especially media services. For example, with respect to movie recommendations on Netflix, Gomez-Uribe and Hunt (2015)[132] confirm a very strong increase in the variety of items that are consumed by users due to the use of personalised recommendations. Furthermore, they highlight the importance of personalisation to persuade consumers to accept and follow recommendations. According to the authors, personalised recommendations also play a significant role in fostering consumer loyalty and reducing subscriber churn. Taken together, they estimate the combined business value of personalisation and recommendations to amount to more than USD 1 Billion per year.

---

[124] Lee, D., & Hosanagar, K. (2019). How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Information Systems Research*, *30*(1), 239-259.
[125] Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159-188.
[126] Lin, Z., Goh, K. Y., & Heng, C. S. (2017). The Demand Effects of Product Recommendation Networks: An Empirical Analysis of Network Diversity and Stability. *MIS Quarterly*, *41*(2), 397-426.
[127] Carmi, E., Oestreicher-Singer, G., Stettner, U., & Sundararajan, A. (2017). Is Oprah Contagious? The Depth of Diffusion of Demand Shocks in a Product Network. *Management Information Systems Quarterly*, *41*(1), 207-221.
[128] Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1-2), 115-153.
[129] Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of retailing*, *80*(2), 159-169.
[130] Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, *29*(1), 84-102.
[131] Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, *24*(4), 956-975.
[132] Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 1-19.

2.2.3 Summary

In e-commerce, retailers collect and process data resources to curate and develop their product portfolio, improve the efficiency of fulfilment services and to facilitate product discovery of customers. To this end, online retailers collect volunteered data and observed behavioural data from users, sales data as well as metadata on their offered products. Whereas product data is either readily available or can be created by investing into the manual or automated annotation and categorization of products, the access to user data is mainly determined by the size of the retailer's user base and its ability to track consumer behaviour inside and outside of its service.

Aggregated sales data is the main input for demand forecasting, which allows retailers to develop their product portfolio according to observed consumer tastes and also to save costs by promoting efficient logistics, optimal warehousing and automated order systems. In this context, empirical studies point to better forecasting performance, when timeseries data on related products' purchase histories are available and predictions can be based on longer timeseries of product sales data. As more fine-granular user data becomes accessible on a large scale, retailers can complement aggregate sales data by usage information, which allows for individual-level targeting of product offers as well as eventually more precise forecasting results.

Operators of marketplaces are in a special position as intermediaries to observe *data on third-party businesses*, especially behavioural data on business-user interactions and purchases from these businesses. Based on this data, the efficiency of the overall marketplace can be improved (e.g., by facilitating the discovery of suitable products), but the access to this data may also give the marketplace operator a competitive advantage in situations, where it competes directly with third-party businesses (e.g., with respect to the development of profitable product portfolios).

Large product catalogues with numerous items per product category, the nuanced differentiation between items, and the availability of a wide set of niche items render *product discovery* a major task for online retailers to convert shoppers into actual buyers. Although electronic markets have drastically reduced the search costs for consumers compared to offline markets, users now often rely on retailers' assistance in dealing with the information overload that they face in these markets. Given the large scale of operations in e-commerce markets, such recommendations can only be supported by automated recommendation systems, which are also regularly found to outperform human advice (see, e.g., Yeomans et al., 2019[133]). To provide users with automated product recommendations, *data on the user base and the product catalogue* are necessary inputs.

To derive personalised recommendations that accurately reflect individuals' interests and preferences, state-of-the-art recommendation algorithms rely on both *explicit feedback data* in the form of volunteered product ratings and *implicit feedback data* in the form of observed user behaviour. As a by-product of user behaviour, implicit feedback data can be collected continuously and at relatively low cost by retailers that already serve an active customer base. Hence, the amount of implicit feedback data that can be collected also scales with the number of active users (see Section 3.1.2.1). Furthermore, retailers must be able to personally identify user and their associated feedback data when deriving personalised recommendations for them.

To overcome the cold-start problem of recommendations systems, collaborative filtering algorithms require a minimum amount of feedback data for each user and also for each product to be considered. *Data on product characteristics* and *user attributes* can help to mitigate the cold-start problem, but contain complementary rather than substitute information to behavioural user feedback data. Therefore, the continuous collection of feedback data is central to gradually improve recommendation performance, as this reduces the sparsity of the user-product data matrix. Specifically, *fine-granular data on user behaviour*, such as data capturing temporal information can increase the accuracy of recommendations. Thus, not only the scale of data collection is relevant for recommendation accuracy, but also the level of granularity at which data can be observed and collected. Moreover, *cross-domain data on user behaviour* from other contexts and services can be used to infer more general preference patterns of individual users and to also identify new similarity relationships across users. Based on this data, recommendation performance can be further improved, especially in the case of new users. Also, cross-domain data, which may be collected from own integrated services

[133] Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403-414.

or, on a less granular basis, from external services such as social networks, can foster the diversity and serendipity or personalised recommendations.

With respect to the role of data for personalised recommendations, it is important to recognise that additional user feedback data (i.e., increasing the depth of an individual user profile) does not only improve the quality of recommendations derived for the respective individual whose data is collected but also exerts a positive externality on the accuracy of recommendations for other users. This is because a deeper user profile allows for better matches when searching for similar users in the process of deriving a recommendation for another user. This positive externality of additional user data can give rise to data-driven feedback effects as discussed in Sections 3.3 and 3.4.

In practice, recommendation systems today are built on a composite architecture that integrates an ensemble of algorithms which can process a variety of data types. In turn, additional data points can contribute to continuous improvements in recommendation quality. With the increasing adoption of machine learning approaches, the processing of fine-granular behavioural data on a large scale is reinforcing this continuous learning paradigm. Moreover, these approaches facilitate the integration of data from different domains and contexts to solve the prediction tasks that are both at the core of personalised recommendations as well as demand forecasting.

In turn, the search for new user data and the need for deeper user profiles may incentivise large online retailers such as Amazon to enter new markets. For example, Amazon's strategy with respect to its home automation devices and voice assistant has been characterised as turning "Alexa into an omnipresent companion that actively shapes and orchestrates your life" (Hao, 2019)[134]. To this end, the firm has recently launched several Alexa products that can be used "on the go" and are thus able to collect user data in many more contexts of users' everyday life. Based on the access to online retailing data resources, in combination with a well-developed computational infrastructure and technical expertise, data-rich e-commerce incumbents may indeed be in an advantageous position to enter other existing or emerging markets.

In the e-commerce market itself, data-driven quality is not the single dimension along which firms compete for consumers. On the one hand, the quality of physical products and the respective product design is to a large extent determined by the engineering capabilities and the innovativeness of the respective product developer and producer. Thus, specialised retailers with high-quality products may attract consumers even if they do not have access to any data that would allow them to forecast demand or derive personalised recommendations. On the other hand, retailers compete for consumers in the price dimension. Whereas in the case of online search engines it is virtually infeasible for competitors to undercut incumbent operators, as the service is already offered at a zero price to end users, retailers in e-commerce can use lower prices to poach customers from incumbents. Of course, this may not be a viable strategy in the long run if it implies continuous financial losses, but it indicates that data on itself is not essential in e-commerce in a narrow sense. Nonetheless, this case study highlights that data plays an important role in establishing competitive advantages within e-commerce markets, which may gradually increase with the access to data resources and thus data can indeed raise entry barriers for new competitors.

2.3 Media platforms and advertising in digital markets

This case study examines the use of data by media platforms, a category that includes a variety of business models, the most prominent of which relies on connecting users and advertisers. This group includes firms that began online and remain online only and others that are the online offerings of publishers or broadcasters that were in the media business pre-internet. Defining this category is significant because some of the major companies operating media platforms rejected being labelled as media for many years in order to maintain a claim on the limited liability for content afforded to services governed by the E-Commerce Directive (ECD) and to position themselves similarly in other jurisdictions.[135] We maintain that two characteristics are key: the service is based on the delivery of content to users and it has some level of responsibility for that content.
Media content can be user generated or professionally produced, and several media platforms carry a mix of both. Regardless of their business model, all media platforms aim to capture the attention

[134] https://www.technologyreview.com/2019/11/05/65069/amazon-alexa-will-run-your-life-data-privacy/
[135] Philip Napoli and Robyn Caplan, 'Why Media Companies Insist They're Not Media Companies, Why They're Wrong, and Why It Matters', *First Monday; Volume 22, Number 5 - 1 May 2017*, 2017, https://doi.org/10.5210/fm.v22i5.7051.

of users. Those media platforms that are essentially offshoots of legacy media or have established themselves from the start as online media have editorial control over their content. Social networks and video sharing platforms that claimed to be passive hosts in the past have since accepted that they too can be considered media. Though the ECD's liability conditions ostensibly remain intact, other EU laws, such as the Copyright Directive and the Audiovisual Media Services Directive (AVMSD), have assigned forms of responsibility to platforms that deliver content, as have EC facilitated self-regulatory mechanisms governing content moderation.[136] The AVMSD, for example, holds video sharing platforms (VSPs) responsible for the organisation of content in the form of hosting, tagging, displaying and sequencing,[137] and therefore places obligations on them to protect minors, combat illegal content, and adhere to common qualitative advertising standards. The category media platforms, therefore, includes the catch-up services of broadcasters, news portals, social media, VSPs, and video on demand (VoD) services. The category does not include platforms that only provide cloud hosting or enable private messaging.

2.3.1 Media platform business models

Data has long been collected from media users, though until the advent of online media they were treated in aggregate as audiences. Even media platforms that do not rely on advertising use such data to maximise their audiences. Though there are hybrids, the business models can largely be broken down into the following models, each of which has different requirements for and ability to collect and use data:

Public Service Media (PSM): This is the model of the free platforms of public service media that do not carry advertiser or require subscription such as the BBC's Iplayer or ZDF's mediathek. Many of these platforms now require registration and collect voluntary and observed user data in order to both engage with user preferences in the fulfilment of their mandate and to generate evidence that can be used to justify their mandate.

Subscription: This model is sometimes also referred to as premium and covers platforms that provide a catalogue or stream of content to users that pay a subscription. These collect voluntary and observed user data and maintain significant sets of non-personal data on the content they carry. Examples of this model are Netflix, Amazon Prime, and Spotify.

Advertising-supported: This model is characterised by a two-sided market with users on one side and advertisers on the other. Even well before the internet, media audiences and advertisers were linked through indirect network effects that were highly dependent on the data that could be gathered on the size and characteristics of the audience.[138] Examples in this category are quite varied in terms of the breadth and depth of data they collect and include news portals like Huffington Post, social networks such as Facebook, and the VoD services of commercial broadcasters. Within this model, an important distinction can be made between those that enable the dissemination of content produced by others and those that engage in the production or procurement of content, but these compete directly in the sale of user attention to advertisers. Voluntary observed and inferred data, as well as non-personal data, are all used in this model, mainly in the delivery of advertising.

Freemium: This model could be considered a hybrid; however, we distinguish it as a separate model here, because it is becoming more prevalent[139] and includes one of the largest media platforms, YouTube, which offers both a free ad-supported option and a subscription ad-free one. In this model, all three types of data are used and again, while both breadth and depth of data are important, services have varying ability to collect it.

Across these models we can identify two main purposes in the collection and use of data: (i) capturing and retaining users, or in other words contributing to the appeal of the platform, and (ii) selling advertising inventory. Key elements of the first purpose are personalisation and service

[136] This refers to mechanisms such as the Code of Conduct on Countering Illegal Hate Speech Online and the Code of Practice on Disinformation. For fuller explanation see Alexandre de Streel and Martin Husovec, *The e-commerce Directive as the cornerstone of the Internal Market: Assessment and options for reform* (European Union, May 2020) https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648797/IPOL_STU(2020)648797_EN.pdf
[137] AVMSD Art 1 para 1(b)
[138] Simon P. Anderson and Bruno Jullien, 'The Advertising-Financed Business Model in Two-Sided Media Markets', in *Handbook of Media Economics*, vol. 1A (Elsevier, 2015).
[139] This is truer in other regions. Asia, for example, seems to have more local and regional freemium media platforms than Europe.

improvement, which have been covered at length in the two other case studies above. So, after a brief discussion of capturing and retaining audiences, we will focus here on the trade in advertising.

2.3.2 Maintaining appeal to users

2.3.2.1 Personalisation

Some degree of personalisation is involved in building the attractiveness of all media platforms, and it runs across web browser interfaces, apps for mobile devices, and often smart TV interfaces. Though they may not depend on monetizing audience attention through advertising or selling subscriptions, just like other services, PSM services also provide personalised recommendations for content and allow users to resume previously played content, both of which require individual user data.[140] Content personalisation can be explicit and implicit, and often it is a combination of both.[141]

Explicit personalisation is based on volunteered information knowingly given for that purpose by the user, such as when a user tells a news app his or her home location to get local news and weather, or when a user sets up both an adult and a child profile in their VoD service's account. This type of personalisation includes the various mechanisms that give users some control over the content they receive or how it is organised. Personalisation can involve giving users the ability to customise the interface or the home page, set filters or alerts, create 'favourites' lists or 'clip' content. The choices users make in the use of such mechanisms contribute to the observed data that informs implicit personalisation.

Implicit personalisation relies on observed data from the individual user and inferences made about their preferences based on that. The core of this type of personalisation is essentially a recommendation for what content to consume. As opposed to linear media services, whose content grids are drawn based on mass-market analysis, a core characteristic of media platforms is an abundance of content and the quality of data-informed ways of helping users navigate this abundance is fundamental to their appeal.[142] Whether it is simply a suggestion of what to watch next, a playlist, or an algorithmically determined news feed, this requires data not just about the users, but also about the content to be provided, the categorisation of which can be an investment in data creation itself. For VoD services, the effectiveness and attractiveness of their recommending are crucial ways they distinguish themselves from the competition.[143] It can include how content is organised and presented in the catalogue, for example, what genres appear and which "new items" are highlighted on the initial catalogue interface, or what "up next" selection appears near the end of a programme. Netflix's approach combines user data with the non-personal data generated by content analysis of all it offers according to 75,000 'microgenres' based on detailed characteristics.[144] Investments into the algorithms that integrated user ratings and other data amounted to millions and thousands of person-hours.[145]

For news media, recommending is often outsourced to third-party "content discovery" companies, the most commonly used of which are Outbrain and Taboola, and it is characterised by the need to account for users' interest in the diversity of content.[146] A person consuming news on a media platform might have a history of choosing content from numerous categories and does not have a clear moment of being finished or having his or her need met, making it very different from personal recommendations in an e-commerce context as we discussed above. It is arguably a deeper and more personal relationship bound by the user's varied, and likely evolving interests, and perhaps the length of their commute or ability to stay awake in the evening.

Matching identifiable personal data to non-personal content data: Content classification is not dependent on the number of users and their data and can be done as part of an initial investment

[140] For PSM platforms and subscription VoD services that operate across smart TVs and other devices a single log in might represent a household of multiple individual users.
[141] Jessica Kunert and Neil Thurman, 'The. Form of Content Personalisation at Mainstream, Transatlantic News Outlets: 2010-2016', *Journalism Practice* 13, no. 7 (2019): 759–80, https://doi.org/10.1080/17512786.2019.1567271.
[142] Gillian Doyle, 'Television and the Development of the Data Economy: Data Analysis, Power and the Public Interest', *International Journal of Digital Television* 9, no. 1 (2018): 53–68.
[143] Xavier Amatriain and Justin Basilico, 'Recommender Systems in Industry: A Netflix Case Study', in *Recommender System Handbook*, ed. Francesco Ricci et al. (New York: Springer Science + Business Media, 2015).
[144] Philip M Napoli, 'Special Issue Introduction: Big Data and Media Management', 2016.
[145] Amatriain and Basilico, 'Recommender Systems in Industry: A Netflix Case Study'.
[146] Yicheng Song, Nachiketa Sahoo, and Elie Ofek, 'When and How to Diversify-A Multicategory Utility Model for Personalized Content Recommendation', *Management Science* 65, no. 8 (August 2019): 3737–57, https://doi.org/10.1287/mnsc.2018.3127.

to drive recommendation based on data volunteered on take-up. However, high quality 45ecommenddation requires tracking the behaviour of each user over time and integrating insight from a breadth of data from other users.

2.3.2.2 Service improvement

Aggregate data enables media platforms to assess the popularity and quality of both the content and features of their interfaces. There has always been a circular relationship between assessing the value, usually as time spent, of content and investment content quality, whether through the mechanism of the two-sided market for advertising media or justifying resources spent on PSM. Now the breadth of data plays a great role, both as an incentive since attracting more users means getting more data, [147] and as an input into ever more sophisticated means of ensuring the provision of attractive content. Observed and volunteered data from users helps platforms improve their services in two important ways. It is now used extensively by content producers and procurers to assess and plan the content they produce or acquire, and by all players to improve their interfaces and functionality. Here the data used is not deep data on individuals, but data that has been aggregated to provide information on audience segments, types of users, and to show tendencies across wider populations of users.

Data on user behaviour is increasingly used to precisely predict what content will appeal to various audience segments but like recommending, this process also relies on automated systems for identifying and classifying content characteristics.[148] Platforms disseminating professionally produced content are investing in the classification of hundreds of characteristics of content, creating "genomes" that are then used to identify trends in user preferences and reactions to particular attributes.[149] The extent of Netflix's content identifying data was mentioned above, and its flagship *House of Cards* series was developed and produced based on data gathered on its users' behaviour and choices in relation to its content.[150] Major broadcasters in developed markets are also investing in this type of precise prediction capability, and a Horizon 2020 project exists to help broadcasters and distributors.[151] Nevertheless, information asymmetries are developing based on the ability to combine user/audience data with content data. The increasing use of data in the management of media, particularly in the production, purchasing and organisation of content[152] will likely have consequences for smaller media and independent producers and distributors.

Matching aggregate behavioural data to non-personal content data: Breadth and freshness of data are important for data-driven service improvement. The data and audience insight that are used to inform decisions about content production and procurement could be shared across multiple firms, if not industry wide.

Media platforms that disseminate UGC and do not invest in the production or acquisition of content invest heavily in the quality of their recommendation systems described above to ensure they have content that will attract users and ensure that users spend time on the platform. Another element of the platforms' appeal is also their implementation of content moderation, including how they handle potentially harmful or illegal content and enforcement of community standards, and the effectiveness of measures aimed at protecting minors. UGC media platforms have come under increasing pressure from policy makers, advertisers and users to improve content moderation and protect users from exposure to illegal and harmful content.[153]

Data from user flagging is a key ingredient in content moderation systems, which are still heavily reliant on human moderators. However, increasingly they rely on *algorithmic commercial content moderation* that identifies content via matching or prediction and then takes an action such as

[147] David S Evans, 'Attention Platforms, the Value of Content, and Public Policy', *Review of Industrial Organization* 2019, no. 54 (2019), https://link.springer.com/content/pdf/10.1007%2Fs11151-019-09681-x.pdf.
[148] Doyle, 'Television and the Development of the Data Economy: Data Analysis, Power and the Public Interest'.
[149] Doyle.
[150] Joseph Turow and Nick Couldry, 'Media as Data Extraction: Towards a New Map of a Transformed Communications Field', *Journal of Communication* 68, no. 2 (2018): 415–23.
[151] ReTV project, 'ReTV-Project', accessed 3 May 2020, https://retv-project.eu/about/.
[152] Napoli, 'Special Issue Introduction: Big Data and Media Management'.
[153] Codes, AVMSD, NetzDG

removal, geo-blocking, account blocking or referral to a human.[154] This relies on databases of content identifiers and getting a match or prediction for a specific piece of content. Perhaps the best known of these is YouTube's Content ID, which is used to identify potentially copyrighted material. Photo DNA, originally produced by Microsoft to combat child pornography, creates sharable databases of *hashes* that can be used to match illegal content.

Since the end of 2016, the Global Internet Forum to Counter Terrorism (GIFCT) has maintained such a common database for identifying terrorist content.[155] Other types of problematic content have more grey areas and even jurisdictional differences. A recent examination of media platforms and computer science literature found slightly differing definitions of hate speech across the platforms, and several different automated approaches to identifying potential hate speech or bullying in content and comments. Media platforms moderating UGC are creating massive proprietary databases that feed into automated means of identifying potentially problematic content in different conditions to implement effective content moderation that will satisfy most users, advertisers and the requirements of regulators.

Harmful and illegal content data: Where there are common standards hash data or other content identifiers, they can be a shared resource.

2.3.3 Selling advertising inventory

The online advertising industry is a complex ecosystem of intertwining relationships with many companies operating at multiple places within it.[156] There have been recent useful attempts to describe the main player's roles and offer simplified illustrations of the process.[157] Here our focus is on the media platforms that sell advertising inventory, which is often represented by sales houses and supply side platforms. It is important to note that while ad serving, data management, verification and measurement are distinct functions some media platforms have these in a house or own properties that fulfil that role.

2.3.3.1 Types of targeted advertising

There is a significant amount of advertising on media platforms that remain non-targeted. Some data is still required in the trading and serving of non-targeted advertising as well, but what interests us most here is the data-intensive trade in targeted advertising on media platforms. There are three main types of targeted advertising all enabled by vast amounts of data and automation, but with varying degrees of reliance on identifiable personal data. These are broad categories that are seeing innovation that sometimes blurs the boundaries, and they are often used in combination.

Contextual advertising is as old as placing airline advertisements in the travel section of a newspaper but has advanced in leaps and bounds with the capacity for automatically generating deep, specific data on content. One of the innovations that helped establish Google's position as the most significant player in online advertising was when, after acquiring Applied Semantics, a text mining start-up, in 2003, it integrated that capacity with its adserving technology to produce a sophisticated tool for enabling precise contextual advertising across any of the websites in the network.[158] Contextual advertising can be done without personal data, such as what the *New York Times* has chosen to do for its online properties, but it can also be combined with personal data for machine learning

---

[154] Robert Gorwa, Reuben Binns, and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society* 7, no. 1 (January 2020): 2053951719897945, https://doi.org/10.1177/2053951719897945.

[155] GIFCT, 'Joint Tech Innovation', *Global Internet Forum to Counter Terrorism* (blog), accessed 3 May 2020, https://www.gifct.org/joint-tech-innovation/. According to Gorwa et al. (2020) The GIFCT database and most others most likely use a perceptual hashing that focuses on characteristics of the content rather than pixels or code so matches do not have to be exact. See those authors for a clear explanation of the technology and categorisation of the types of content ids.

[156] Sally Broughton Micova and Sabine Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing' (Centre on Regulation in Europe (CERRE), April 2019).

[157] Competition & Markets Authority (CMA), 'Online Platforms and Digital Advertising Market Study:Observations on the CMA's Interim Report', 18 December 2019, https://assets.publishing.service.gov.uk/media/5dfa0580ed915d0933009761/Interim_report.pdf; Stephen Adshead et al., 'Online Advertising in the UK' (London: UK Department of Media Culture and Sport, January 2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/777996/Plum_DCMS_Online_Advertising_in_the_UK.pdf.

[158] Martin Moore, *Democracy Hacked:Political Turmoil and Information Warfare in the Digital Age* (London: Oneworld, 2018).

enhanced targeting.[159] Audiovisual media services have developed algorithms that can identify features in their content and match advertisements with the content for their linear and VoD services, so that for example if a in a scene in a sitcom one character happens to be playing a video game when the others walk into the room arguing, there would be an ad for Playstation or Nintendo in the next ad break.[160]

The *segment-based* type of advertising most resembles how advertising was done in the linear broadcasting world, which operates similarly using panel surveys and audience measurement data. Audience segments of homogenous groups are created based on demographic characteristics, social media use habits, cognitive styles and affinity to celebrities or other taste preferences.[161] This type of advertising relies on the insight generated from broad pseudonymised or anonymised and aggregated data from a variety of sources. In the end, serving an advertisement to an individual user requires matching that individual with that audience segment, so their personally identifiable data, sufficient to put a user into the desired segment (or exclude them), is required.

The newest type of targeted advertising is *behavioural* advertising, which delivers advertising to specific users based on deep personal data about them and their previous activities online. This data on an individual's characteristics and past tendencies are used to predict potential their likely future action and thus the potential effect of an ad.[162] Re-targeting falls into this category and involves serving ads to users identified, often through their clickstreams data, as previous customers, visitors to the website, or as having searched for the product or service.[163] This requires generating detailed user profiles based on very deep data sets with identifiable individual users.

2.3.3.2 Individual User Profile Data

Both segment-based and behavioural advertising are reliant on individual user profiles. The data used to build and maintain user profiles for advertising is not given at a singular point in time but is continually refreshed by the observance of user behaviour,[164] and sometimes this observance is even in real time, such as in the conditional sequencing of ads.[165] The ability to access and sync data from multiple sources that often use different IDs gives some players in the ecosystem much greater capacity to generate granular and continuous profiles than others.

Advertisers and media platforms have data, gained directly from their users and customers. This *first-party data* comes from the registered use of their websites, from loyalty programmes, apps, and a variety of other interactions online and, for some advertisers, offline.[166] Global companies with multiple services such as Google and Facebook have deeper and broader pools of first-party data than others due to the vast numbers of individual users and interaction points from which they can draw data. Those collecting first-party data are responsible for gaining consent for its processing and then protecting that data on behalf of the individual. Nevertheless, consent usually covers all the sharing of this data required for advertising. For example, an automotive brand may have first-party data on those who have visited its website, those who have been for a test drive, and previous buyers, and this data will be used by their media agency, and a demand side platform at least and possibly an ad exchange and others as well. Companies use data management platforms to keep track and can outsource the maintenance and utilization of user profiles.

Although consent is less straightforward, *third-party data* also features heavily in much of the individual user profiling that drives segment-based and behavioural advertising. Research tracking third-parties' data collection shows that those related to advertising vastly outnumber any other kind of third-party data collectors and that overall the number has changed little despite the

[159] Forbrucker Radet, 'Out of Control: How Consumers Are Exploited by the Online Advertising Industry', 14 January 2020, https://www.forbrukerradet.no/undersokelse/no-undersokelsekategori/report-out-of-control/.
[160] Broughton Micova and Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing'.
[161] Victoria Fast, Daniel Schnurr, and Michael Wohlfarth, 'Data-Driven Market Power: An Overview of Economic Benefits and Competitive Advantages from Big Data Use', *Available at SSRN 3427087*, 2019.
[162] Sophie C. Boerman, Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius, 'Online Behavioral Advertising: A Literature Review and Research Agenda.', *Journal of Advertising* 46, no. 3 (July 2017): 363.
[163] Fast, Schnurr, and Wohlfarth, 'Data-Driven Market Power: An Overview of Economic Benefits and Competitive Advantages from Big Data Use'.
[164] Jan Kraemer and Michael Wohlfarth, 'Market Power, Regulatory Convergence, and the Role of Data Indigital Markets', *Telecommunications Policy* 42 (2018): 154–71.
[165] This refers to when the choice of which version of an ad is served to a user next depends on some kind of interaction with the previous ad, such as with a dependent narrative. For accounts from practitioners see Broughton Micova and Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing'.
[166] For an account of how offline retailers collect first party customer data see Joseph Turow, *The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power* (Yale University Press, 2017).

implementation of GDPR,[167] though some decline in third-party presence on news providers' pages has been observed.[168] Most of these third-parties have very small footprints and studies consistently show that the most prevalent by far are those owned by Google and Facebook with others in the top echelon being Amazon, Oath (Verizon), Adnexus, Criteo, Adform, Rubicon, TMRG and Twitter, and Comscore.[169]

Much of the data gathered by third-parties is done by using cookies that allow them to collect data on online behaviour connected to an identifiable individual. A unique identifier, most often a cookie, allows whoever manages it to find the user again across other sites and services. Whereas first-party cookies are those managed by the owner of the website the user is consulting, third-party cookies relate to other servers called by the page that have been allowed to make that connection by the first party.[170] The first party owners of the interface with the user, the website, must manage to get consent for the various third-parties whose servers will be allowed to identify and track that user. Advertising supported media have to enable the third-parties that are used by advertisers and agencies to make sure their inventory is recognised by their demand side platforms or otherwise identified as reaching the desired user and purchased.

Third-parties also engage in various types of "fingerprinting" especially when their cookies are not allowed. This arguably more invasive process involves collecting information unique to a user's device to establish an identity and can include the combination of plugins installed, browser and other settings, and even details of the cache or which fonts have been installed. Fingerprinting can be done across devices and can be used to continue tracking users that have taken steps to opt-out.[171] Tracking individual users on mobile devices that are not using browsers is enabled by unique IDs such as the Android Advertiser ID. Though users can 'reset' their Android Advertiser ID and similarly clear other identifiers, if the tracker has another identifier that has not changed it can immediately re-establish the connection to the individual users. Testing done by the Norwegian Consumer Council found static IP address was being received by trackers along with OS-based persistent IDs and thus could be used to recreate profiles.[172]

The collection of the identifiable personal data that feeds into individual profiles is changing because of a privacy motivated push-back on tracking, particularly with regard to third-parties. All the major web browsers have now moved towards blocking third-party cookies. Safari has long inhibited third-party cookies, Mozilla and Chrome have both announced measures to stop various ways of fingerprinting, but only Chrome has put forward an alternative to the current cookie-dependent system.[173] Its Privacy Sandbox is a set of initiatives, and the model it is establishing is one in which the necessary processing of data is done on the device at the browser level so that it can be used for measurement or targeting without the individual being identified or the data being lifted.[174] The Sandbox, and IAB's recently launched REARC initiative, also attempt to replace cookies and mobile IDs with privacy-centred solution.

[167] Jannick Sørensen and Sokol Kosta, 'Before and after Gdpr: The Changes in Third Party Presence at Public and Private European Websites', 2019, 1590–1600.
[168] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielson, 'Changes in Third-Party Content on European News Websites after GDPR' (Oxford: Reuters Institute for the Study of Journalism, August 2018), https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-08/Changes%20in%20Third-Party%20Content%20on%20European%20News%20Websites%20after%20GDPR_0_0.pdf.
[169] Sørensen and Kosta, 'Before and after Gdpr: The Changes in Third Party Presence at Public and Private European Websites'; Libert, Graves, and Kleis Nielson, 'Changes in Third-Party Content on European News Websites after GDPR'.
[170] Kevin Mellet and Thomas Beauvisage, 'Cookie Monsters. Anatomy of a Digital Market Infrastructure', *Consumption Markets & Culture* 23, no. 2 (2020): 112.
[171] Forbrucker Radet, 'Out of Control: How Consumers Are Exploited by the Online Advertising Industry'.
[172] Forbrucker Radet.
[173] Mozilla, 'Security/Anti Tracking Policy', Mozilla Wiki, 9 July 2019, https://wiki.mozilla.org/Security/Anti_tracking_policy.
[174] Sam Tingleff, 'Explaining the Privacy Sandbox Explainer', *IAB Tech Lab* (blog), 27 March 2020, https://iabtechlab.com/blog/explaining-the-privacy-sandbox-explainers/.

Identifiable personal data: The ability to utilize deep, often continual, data about specific individuals is a key differentiator among firms in the business of targeting using individual profiles. Access to (and consent to use) first-party data is a valuable asset, as the ability of third parties to collect and use personal data is not stable. Media platforms selling inventory have an incentive or, arguably, are compelled to allow third-party data collection on their properties to ensure their inventory is recognized by advertiser's buying systems.

2.3.3.3 Aggregate user data and insight

The personally identifiable data gathered through the various means mentioned above is used in the aggregate for the insight upon which segment-based advertising is conducted. The IAB has recently set out an *audience taxonomy* of nearly 1700 demographic, interest and purchase intent characteristics to bring some standardisation to audience segmentation.[175] Previously, actors dealing with segment-based advertising have established their taxonomy. While established segments may be relatively static, predicting the behaviour of people in any given segment in response to an ad requires the insight from a breadth of observed and volunteered data from individuals belonging to that segment and this data needs to be fresh and updated, if not continuous. The data in question consists of highly detailed individual, though pseudonymised, user profiles that have been compiled by syncing user identifiers across several data sources. These sources often include real-time behavioural data and regularly updated data from offline sources, which poses challenges in terms of data quality, reconciliation or syncing, consistency and heterogeneity of the schema used in aggregation.[176]

Advertisers will aim to reach specific audiences depending on their objectives, which could be to reach new groups of customers defined by certain characteristics or to reach additional customers who are similar to those that are already their customers. The insight into how to reach customers falling into those segments comes from pseudonymised or anonymised aggregate data. A host of third-parties exist that specialise in analysing this kind of data. Several media agencies and specialist companies, such as DunnHumby and Criteo, incorporate both online and offline on behalf of advertisers to predict how particular segments will behave, understand how best to reach them, and create "look alike" categories that are fed back into media platforms to target potential customers. Media platforms selling inventory, in turn, must have sufficient data on their users to present them as audience segments. Data management platforms are also either engaged by or have been created by media platforms selling ad inventory to help them turn their user data into insight that feeds into segment-based targeting.

Reaching the individual users that then fall into any given segment requires identifiable personal data. The often highly granular segment characteristics must be joined up, in real-time, with identifiable personal data to determine segment membership. The matching of users to segments for ad delivery must be done within a consent-managed environment, so by the media platform's in-house or partner data management platform or otherwise by the media platform selling the inventory that had acquired the consent of the individual user. Facebook, for example, offers tools that draw on the great breadth of aggregate data it has for planning segments and then enables delivery to its users, but one cannot extract the data connecting an individual user with that segment to reach the user in another way. Criteo attempts to mimic this capacity in the open web by drawing on first-party data from retailers and the consent it has from inventory holders that allows it to identify individuals within segments.

Aggregate 'audience' data: Organisations dealing in segment-based advertising will compete on the granularity and reliability of their segments. Both demand and supply side actors must invest in the collection of deep personal data that is aggregated to establish segments.

2.3.3.4 Campaign data for measuring effectiveness and efficiency

There is overlap between the data used to gain insight into audiences and data which demonstrates the effectiveness of a campaign. For both individual behaviour data in the form of clicks and conversion, and indicators of attention such as the duration of views, are aggregated. Instead of

[175] IAB Tech Lab, 'Audience Taxonomy 1.1', *IAB Tech Lab* (blog), April 2020, https://iabtechlab.com/standards/audience-taxonomy/.
[176] Hazem Elmeleegy et al., 'Overview of Turn Data Management Platform for Digital Advertising', *Proceedings of the VLDB Endowment* 6, no. 11 (2013): 1138–49.

being aggregated to indicate characteristics of or make predictions for an audience segment, it is aggregated according to specific advertising campaigns. In addition to basic impressions which indicate reach, click through rates (CTR), conversion rates (CVR), and other post-exposure behaviour metrics are tracked for each campaign and perhaps distinct elements of it. Much of this data has been cookie-dependent either from first-party inventory holders tracking their own users' behaviour in response to the ad they carry, or from third-parties that offer verification, measurement and attribution tracing.

This data feeds into the econometric modelling done within agencies for planning purposes, and data to which media platforms have access is further analysed and presented to demonstrate the effectiveness of their inventory.[177] The data, therefore, informs price setting in direct buys and wider assessments of the value of various media platforms and types of inventory options. Crucially, it also informs bidding strategies in programmatic buying as it serves to predict the value of the inventory up for bid.

In the programmatic auction-based trading DSPs' bidding strategies are based on highly complex prediction based on data from past campaigns. Computer and data scientists continue to devise new ways to improve the ways these predictions are made. As Lai et al. describe there are three main elements to DSP strategies: value evaluation based mainly on past CTR and CVR, price prediction based on the past winning bid, and budget control through pacing bidding. The way advertisers or their agencies ensure clicks and conversions happen is through finding the conditional probability of clicks or conversions given feature combinations of audiences, platforms and specific ads.[178] Those whose inventory is up for bidding necessarily have a stake in how these calculations are made and what data feeds into them.

In today's media environment users are often served the same ad or ads from the same campaign for the same product or service through multiple platforms and devices, therefore attributing any resulting action to the right place is a science unto itself. Sometimes simply the last clicked ad or the first ad to make an impression are attributed the credit for the action the user takes. In other models, the first and last receive an equal and larger share, while the rest is split among all other locations where the user was served relevant advertising. Data-driven attribution (DDA) involves complex algorithms that distribute attribution across all the user's exposure point.[179] DDA requires deep, real time, personal data as conversion or click event data are integrated with user behaviour tracking data. To use DDA, a company must be able to follow a unique identifier across all advertising exposures to collect clean reliable data, quickly or even in real time as it is used to train the machines conducting real time bidding in programmatic systems.[180]

Campaign data: Being able to predict as accurately as possible the effect of an ad is fundamental to establishing its value both for demand and supply side actors. For media platforms, access to data generated by user interaction with the advertising around their content across all delivery systems through which it reaches an audience is necessary for them to be able to engage on relatively even terms with demand side actors.

2.3.3.5 Transaction data

Finally, the conduct of all three types of advertising is guided by data tied to the actual trade of the inventory, including data tied to each transaction. Cost per impression (CPM) has long been a measure of value in advertising, assessed, for example, in broadcasting based on audience measurement data and the price paid for the inventory. For online media platforms, a much greater number of indicators of value are recorded, many linked to the campaign data discussed above. In the case of direct buys, the price per impression can be accompanied by details on the quality of the

[177] Broughton Micova and Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing'.
[178] H. Lai et al., 'Predicting Traffic of Online Advertising in Real-Time Bidding Systems from Perspective of Demand-Side Platforms', in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, 3492, https://doi.org/10.1109/BigData.2016.7841012.
[179] Kyra Singh et al., 'Attribution Model Evaluation', 2018; Eustache Diemert et al., 'Attribution Modeling Increases Efficiency of Bidding in Display Advertising', in *Proceedings of the ADKDD'17*, 2017, 1–6.
[180] Diemert et al., 'Attribution Modeling Increases Efficiency of Bidding in Display Advertising'; Richard Parboo, 'How to Optimise Your Marketing Attribution', *Iab UK* (blog), 9 July 2019, https://www.iabuk.com/opinions/how-optimise-your-marketing-attribution.

impression, such as placement, viewability and sound status, if relevant. In combination with the CTRs and CVRs mentioned above, cost per click or conversion can be calculated.

With the advent of programmatic real-time trading, there are billions of individual transactions being recorded. Data is produced on the prices paid, the winning bidder's offer (as in second price auctions these are not the same), any floor set by the seller and it is also produced about the conditions of the ad placements, for example on placement, whether sound is on or off if relevant, viewability metric. As with the attribution models mentioned above, this data is used to train for future bidding and contributes to the inherent information asymmetry in trading through ad exchanges that advantages demand-side players.[181] Media platforms offering inventory on the open web are increasingly turning towards header bidding, which allows multiple players on the demand side to bid for ad placement. This not only gives them the chance to get higher value for their inventory because by offering it directly to many supply-side platforms or exchanges at once, rather than accepting the valuation of only one but also gives them the chance to gather data from each of those interactions, which can inform their choices about floor costs and selling strategies in the future.

Transaction data: Because of the complex strategies involved in buying and selling, especially through real-time bidding, a breadth of detailed transaction data, which includes both non-personal financial and contextual data on vast numbers of individual trades, is needed by both demand and supply-side actors.

2.3.4 Economic value creation

Creating value for advertisers is directly related to the value that advertising options have for media platforms. The effectiveness of targeting is now well supported, and there is evidence that it may be more so with greater reliance on first-party data within trusted environments since trust and transparency increase effectiveness.[182] However, the value of behavioural advertising for different types of competing firms that have access to the same targeting capability is less straightforward. Chen and Stallaert emphasize that "while small advertisers are generally better off under behavioural targeting by winning their targeted users, the dominant advertiser might or might not be better off. The dominant advertiser is worse off under behavioural targeting when he has a significant competitive advantage over his competitors because, under traditional advertising, he would otherwise grab a larger group of users and still realise a decent payoff."[183] This indicates that there will remain a market for less granularly targeted and non-targeted advertising, especially for big commercial brands, such as fast-moving consumer goods, retailers, and basic services.

It has been argued that multi-homing advertisers will easily choose smaller, cheaper per-click options in an online environment, and therefore data at great breadth and depth is not essential.[184] However, advertisers do not multi-home based on cost per click.[185] Instead, they essentially spread their bets based on complex models that attempt to predict the effect or return on various advertising options. The trade in advertising is based significantly on prediction and this depends on access to a pipe stream of deep and broad data. Many companies exist that manage the data and provide predictive insight mainly on the demand side. These have been heavily reliant on cookies so far, but at the start of 2020 Chrome joined Mozilla and Safari in disallowing third-party cookies and taking measure to inhibit fingerprinting. With browsers moving away from third-party cookies, first-party data such as from advertisers themselves and through companies like Criteo becomes even more crucial and valuable.[186]

---

[181] Alison Schiff, 'Can LiveRamp Survive The Cookie Apocalypse?', online magazine, *Adexchange* (blog), 9 March 2020, https://www.adexchanger.com/data-exchanges/can-liveramp-survive-the-cookie-apocalypse/.
[182] Tami Kim, Kate Barasz, and Leslie K John, 'Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness', *Journal of Consumer Research* 45, no. 5 (2019): 906–32.
[183] Jianqing Chen and Jan Stallaert, 'An Economic Analysis of Online Advertising Using Behavioral Targeting', *Mis Quarterly* 38, no. 2 (2014): 446.
[184] Andres V Lerner, 'The Role of'big Data'in Online Platform Competition', *Available at SSRN 2482780*, 2014.
[185] Patrick Barwise and Leo Watkins, 'The Evolution of Digital Dominance: How and Why We Got to GAFA', in *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, ed. Martin Moore and Damian Tambini (Oxford: Oxford University Press, 2018); Broughton Micova and Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing'.
[186] Anna Hensel, 'Cookiepocalypse: What the Death of the Third-Party Cookie Means for Retailers', online magazine, *ModernRetail* (blog), 20 January 2020, https://www.modernretail.co/platforms/cookiepocalypse-what-the-death-of-the-third-party-cookie-means-for-retailers/; Schiff, 'Can LiveRamp Survive The Cookie Apocalypse?'

To coincide the changes to Chrome, Google announced its Privacy Sandbox. Some have already raised concerns that Chrome's move to stop third-party cookies and replace them with Sandbox may result in further strengthening of the position of Google and Facebook that are rich in first-party data.[187] Mellet and Beauvisage also warned that if third-party data collection is limited, "the Facebook consumer data platform appears as the main pretender to the succession of the cookie-based infrastructure"[188] implying that tools for using the open web will necessarily be replaced by in-ecosystem tools for utilizing first-party data held within those ecosystems. The implementation of the GDPR has already been shown to have reduced the number of third-party data vendors and led to increased concentration among ad-tech providers.[189] The changing policies towards cookies are still materialising and the consequences thus remain to be seen.

Some larger audiovisual media services that operate media platforms are essentially pooling their datasets by cooperating on tools to offer addressable and programmatic buying options at scale both on a European level, for example through the European Broadcasting Exchange[190], or within national jurisdiction, such the cooperation between RTL and ProSiebenSat1.[191] These can all be seen as attempts to match as much as possible the capacity to provide audience segments at granularity and scale and to capture the campaign and transaction data needed to assess and attest the value of their inventory. When they disseminate their content through other media platforms, they may be able to sell the advertising around it themselves and/or receive a share in revenues, but they do not receive the equivalent of the campaign data concerning those ads.[192]

Campaign and transaction data is not just important for designing strategies for approaching real-time bidding and setting prices for direct buys. At scale, it is also essential for demonstrating effectiveness and efficiency to advertisers. Experiments aimed at assessing the effectiveness of advertising on media platforms have given vastly different indications as to how much data is needed, but 'lesser' options involved tracking three million users over two weeks.[193] Experiments may be useful for specific queries, and some demand-side tools enable these to be run within campaigns to aid planning and buying.[194] The advertising business, however, depends on a continual flow of data into advertiser key performance indicators (KPIs), which are largely derived from the integration of campaign and transaction data, so non-personal and aggregated personal observed and inferred data. Given the increasing focus on demonstrable efficiency among advertisers,[195] inventory of a media platform must fare well in relation to not only the traditional CPM, but also cost per view or full view, per click, per conversion, per purchase and variety of ways of defining cost per action. In a sense, this may give rise to a cold-start problem for new media and content offerings due to missing data, similar to the cold-start problem for new products in the case of personalised recommendations in e-commerce (see Section 2.2).

Contextual advertising requires extensive investment in classifying content and in systems that match advertising to that content. The IAB has developed a context taxonomy, which can help in this regard. It remains an option much less dependent on personal data. A new entrant might start in the market with contextual inventory until their users generated enough scale and duration in the collection of first-party data and the establishment of ties to third-parties required for segment-based or behavioural advertising. However, campaign and transaction data for establishing value and evidencing effectiveness will be equally necessary.

---

[187] Damien Geradin and Dimitrios Katsifis, 'Online Platforms and Digital Advertising Market Study Observations on CMA's Interim Report', February 2020, https://assets.publishing.service.gov.uk/media/5e8c8a4b86650c18c6afeab5/200212_Prof._Damien_Geradin_and_Dimitrios_Katsifis_Response_to_Interim_Report.pdf.
[188] Mellet and Beauvisage, 'Cookie Monsters. Anatomy of a Digital Market Infrastructure', 19. Here they refer to
[189] Garrett Johnson and Scott Shriver, 'Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR', *Available at SSRN*, 2019.
[190] EBX, 'European Media Corporations Agree on Joint Venture', *European Broadcasting Exchange* (blog), 11 September 2017, http://ebx.tv/?page_id=269.
[191] Chris Dziadul, 'RTL and ProSiebenSat.1 Ink Addressable TV Joint Venture', *Broadband TV News*, 5 June 2019, Europe edition, https://www.broadbandtvnews.com/2019/06/05/prosiebensat-1-rtl-ink-addressable-tv-joint-venture/.
[192] Demand side actors using, for example, Google's Ad Manager 360 can access and extract aggregate campaign data through Google's Data Transfer system and Facebook offers similar through its Facebook Ads Manager for Excel function.
[193] Garrett A Johnson, Randall A Lewis, and David H Reiley, 'When Less Is More: Data and Power in Advertising Experiments', *Marketing Science* 36, no. 1 (2017): 43–53.
[194] Facebook, for example, offers A/B split testing within its Ads Manager (see https://www.facebook.com/fmp/agencies/measurement), though without the control groups used by Johnson et al. (n 58)
[195] Broughton Micova and Jacques, 'The Playing Field for Audiovisual Advertising: What Does It Look like and Who Is Playing'.

In the end, media platforms differentiate based on the quality of their targeting, which largely determined by the data that can be connected to their users. Strength depends not only on the breadth and depth of a media platform's first-party data but also on the quality and nature of their connections to others in the ecosystem such as DMPs and third-parties. These determine how easy the platform makes it for advertisers to target and track effectiveness.

As advertising-dependent media platforms compete fiercely with each other for advertising expenditure, they are also competing with media platforms based on other business models for the attention of users. The popularity and extent of PSM platforms vary drastically from country to country, but subscription media platforms are expanding in reach nearly everywhere. In the subscription model, there are not the same network effects as for those in the two-sided market in which advertising-supported media operate. Though their recommender systems also rely on personally identifiable data, much of it is volunteered in the form of ratings or reviews and profile information, and the collection of observed data required for recommender systems is likely more shallow than the kind of tracking upon which so much advertising depends. The initial strength of a media platform's recommender system is likely outweighed by the quality of their content, so the ability to be able to invest in the production or procuring of high-value content is more important. Increasing diversification in this area has been driven by companies that already have rights to a wealth of content of large amounts of capital to invest. Netflix, Amazon Prime and recently Disney+ not only draw user attention but also may shape their expectations.

Users have a choice between a monthly subscription, some public media, and being targeted by advertising. The pure advertising model, in general, may be in decline in online media. Many press publishers maintained a combination of subscription and advertising in their online offering. For user-generated content, the future might be more freemium models, such as YouTube Premium, that seem to offer users a choice between paying with their attention (for advertising) and paying with money. Such services will keep the attention of users that do not want the intrusiveness of advertising, yet will still gather a lot of their data, and may procure data generated from their response to advertising (views, clicks, etc.) on other services. However, the more users they have for both free and subscription version the better their recommendation systems and therefore the more attractive they are, and the data from subscribers can still be aggregated to inform audience segment insight or even to profile those individuals for targeting on other services.

03

THE ECONOMIC VALUE OF DATA

3 The economic value of data

The value of data is derived from the information and knowledge that can be extracted. Therefore, the economic value that can be created from data is necessarily context-dependent. At the same time, the same data may be the basis for value creation in very different contexts. Thus, while the economic benefits of data need to be evaluated with respect to the specific use cases, the presented case studies have indicated that different types of data and dimensions of value creation can be categorised and differentiated when characterizing the role of data in digital markets.

The three case studies show that data in digital markets has become a key input resource for businesses instead of only a factor that marginally improves products and processes. Still, about search, media and e-commerce markets, firms will likely be able to provide a basic service without data or with data that is publicly available and can be collected (e.g. the search index data or content characteristics), although this may require large investments. However, as shown for all case studies, data is important to continuously improve the quality of service. That is mainly because prediction tasks are now at the heart of most digital business models (Calvano & Polo, 2020). More precisely, businesses are mostly concerned with predicting users' needs, their preferences and their behaviours, both at an aggregate and an individual level. Specifically, this allows firms to offer more relevant search results, more accurate demand forecasts, more interesting predictions or more effective advertising. By doing so, businesses can significantly improve quality and create economic benefits in various ways and along different economic value dimensions. For example, as highlighted in the case study on e-commerce, more accurate recommendations may facilitate the discovery of niche content and allow a retailer to serve a wider user base with heterogeneous interests, but also increase customer satisfaction, which lowers customer churn. Taken together, a basic service that does not benefit from data-driven quality improvements will often be insufficient to attract users and to grow a viable customer base. In consequence, the economic benefits may provide data-rich firms with long-term competitive advantages, even though the data may not be essential per se.

In this section, we draw on the specific insights from the previous case studies and additional literature to characterize the criteria that determine the economic value and potential competitive advantages from data in more general terms. Specifically, whether and when data may establish sustained competitive advantages in digital markets and what implications to draw for the contestability of the respective markets is determined by (i) whether and to which degree other firms can duplicate data resources of incumbents and (ii) how the economic value from data is related to the scale and quality of the data resources.

Competitive advantages based on the use of data may only be transitory if competitors and entrants can gain access to the same data or substitute data from which the same insights can be inferred. Hence, the conditions under which duplication of data is feasible and more precisely the barriers to collect data determine whether digital markets may remain contestable when data delivers significant economic value. To this end, the (non-)rivalry in data collection, firms' ability to collect first-party and third-party data, as well as the access to external data sources, are important criteria to consider.

Scale advantages in data collection and data processing may establish long-term competitive advantages for incumbents and raise entry barriers, thus endangering contestability of markets that are dominated by data-rich incumbents. Whereas the prior policy debate has often centred around a general notion of scale or volume of data, we propose to differentiate between two basic dimensions of input data sets in digital markets (as illustrated in Figure 2). Firstly, we highlight that scale may refer to the *breadth* of a data set, which measures the number of data points that are available for a specific item, i.e., for example, a product or a specific search query. Hence, this dimension increases with the number of individuals from whom data can be collected. Secondly, the scale may refer to the *depth* of a data set, which describes how much data is available for a specific individual. Data grows along this dimension if information about an individual is collected over time and possibly across domains and services. If a new data point is collected that is traceable and thus the data can be associated with an individual, this data contributes to both dimensions. In contrast, anonymous data collection adds to the breadth of data, but cannot increase the depth of a data set. As highlighted in the following this has important ramifications on what competitive advantages we expect from data collection. Moreover, it also has implications for what can be achieved by sharing different types of data as a potential remedy (see Section 5).

Economic value and competitive advantages from data do also depend on the *quality of data*. In particular, timeliness of data may represent a key requirement for some services and fresh data is likely to improve service quality for most prediction tasks. Moreover, the accuracy of data, which also significantly affects the economic value that can be derived, is determined by the type of data collection that a firm can undertake. Finally, it is important to consider the granularity or level of detail that is contained within a data set.

Taken together, scale or quality advantages and initially superior access to data may give rise to feedback effects, such that data-driven competitive advantages are magnified over time as improved service quality from data leads to more users and this then turns into access to even larger data sets. Data-driven network effects and economies of scope in data may not only protect incumbents in their core markets but also facilitate expansion into adjacent markets if they can leverage existing data resources.

Finally, the economic value from data also depends on complementary inputs such as the technological and physical infrastructure, developers and data scientists, and the algorithms for data processing. Supply-side economies of scale and scope, in addition to competitive advantages, may be obtained from the access to data. We review and discuss these issues in the following.

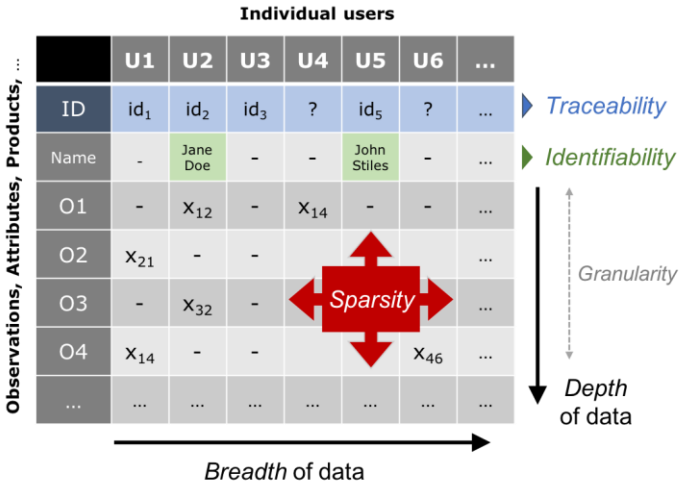

Figure 2: Illustration of data dimensions that determine the economic value of data.

3.1 Duplication of data resources

The presented case studies demonstrate that user data is one of the central inputs for the generation of economic benefits and specifically the continuous improvement of service quality. Thus, firms' ability to duplicate an incumbent's data assets or to gain access to alternative data resources is key for competition. To this end, firms can collect user data as volunteered or as observed data or buy data from third-parties as highlighted in the individual case studies. With respect to whether competitors can effectively duplicate these data sources, there has been a lively debate on the degree of rivalry in data collection from consumers and the ability of incumbents to exclude other firms from access to data resources.

3.1.1 Data collection from consumers

In principle, data is *non-rivalrous*, which means that the same data can be shared and collected by different entities without depleting the source of the data or reducing the availability of data for others. This characteristic has been emphasized by some scholars who claim that data is ubiquitous, as consumers are willing to share their data over and over again with different services, frequently multi-home similar services, and that specialized data brokers make data available to everyone who wants to buy it (see, for example, Lambrecht and Tucker, 2015[196], and Tucker, 2019[197]).

[196] Lambrecht, A. and Tucker, C.. Can Big Data Protect a Firm from Competition? (Dec. 18, 2015). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2705530
[197] Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. Review of Industrial Organization, 54(4), 683-694.

This is contrasted by the empirical findings that – despite the multitude and variety of websites and online services available – consumers' attention is highly concentrated on a few sites and even fewer firms.[198] Because most of the relevant data inputs are generated as a by-product of consumers' usage, firms that can directly observe this usage are in a superior position to collect data on user behaviour at an individual level and at large scale. Although it is generally conceivable that firms may collect user data also outside of their service directly from consumers, this data will usually be of lower quality and it will put firms at a continuing cost disadvantage if data outdates quickly (see Section 3.5). For example, firms could pay internet users to install browser tracking extensions or other traffic monitoring apps, which record users' web browsing behaviour (see, e.g., the Nielsen mobile panel[199] and the browser offered by Cocoon[200]). However, only a small share of users will agree to such an explicit, transparent collection of their data due to privacy concerns and even participating users may switch to alternative browsers or turn off tracking in specific situations, e.g., when entering sensitive search queries. In consequence, this will lead to incomplete and biased data samples with low accuracy, which depletes the value of the data.

3.1.2 First-party and third-party data collection

Digital web tracking technologies make it possible to collect data on consumers' behaviour directly on the server side. As illustrated in Section 2.3 for the online advertising of media platforms, *first-party data* refers to data that is collected on a firm's website, application or device, while *third-party data* is collected by tracking users and observing their behaviour outside of the firms' properties.

3.1.2.1 Scale and incumbency advantages in first-party data collection

Because consumers' usage of digital services is inherently intertwined with the creation of data, first-party data collection is prevalent in digital environments. As firms with an established user base and existing customer relationships can collect broader data sets more quickly and at lower costs, there are substantial *scale advantages* in first-party data collection. On the one hand, firms can use established relationships with users to contact and directly reach out to them to collect (volunteered) explicit feedback data, e.g., in the form of ratings or reviews. On the other hand, firms can collect (observed) fine-grained data on users' transactions and interactions carried out on their services. Thus, for example, user-generated product ratings or observed data on search queries accumulate more quickly with a larger customer base and when more users interact with a firm's service.

Moreover, established firms are likely to have an advantage when personal data is collected. Several empirical studies have found that the *trust of consumers*, which is, e.g., built by a firms' brand image, plays a significant role in accepting data-driven services and in mitigating consumers' privacy concerns, which inhibit the willingness to share data with a firm.[201] In consequence, entrants, unknown to users, may find it especially challenging to persuade consumers to share extensive sets of personal data with them.

3.1.2.2 Creation of individual user profiles by tracing users within and across services

Furthermore, first-party data collection facilitates the *creation of user profiles*. As users frequently register and authenticate themselves with a unique user account when using a digital service, their behaviour can be traced over time and data can be merged across website visits or product purchases. In many cases, users will also enter personal information, such as their name and address, when registering a user account, which renders the entire associated user profile personally identifiable. This may also facilitate the matching with additional data from other services of the firm or with data from external sources. As accurate user information is often required for the service itself (e.g., the correct name and address are required for shipping and billing in e-commerce), data quality of the volunteered personal information is usually high. Even if users do not authenticate themselves with a user account, their behaviour can easily be tracked and traced across sessions by

---

[198] For example, the European Commission found in the context of the Google AdSense case that Google had a market share of generally over 90% in 2016 the market for general search in all Member States. See https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770.
[199] https://mobilepanel2.nielsen.com/
[200] https://www.onavo.com
[201] See, e.g., Bleier, A., & Eisenbeiss, M. (2015). The importance of trust for personalised online advertising. *Journal of Retailing*, *91*(3), 390-409 and Wang, W., & Benbasat, I. (2005). Trust in and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3), 72-101.

browser cookies or by passive fingerprinting techniques, which identify revisiting users based on the unique configuration profiles of their browsers and devices.

A small number of companies has been successful in establishing *integrated services ecosystems* that now span across multiple service domains, markets and layers of the digital value chain, which allows these firms to collect very deep data sets. By gaining a holistic view on users' global behaviour and recording user behaviour across services, they can create and enrich user profiles based on first-party data that span interests and preferences beyond a single domain. Users of services that are part of larger services ecosystems are often asked to give their general consent for data collection across all services. Hence, the collection of personally identifiable data may also be facilitated for these integrated firms, whereas competitors in individual markets must obtain purpose-specific consent.

3.1.2.3 Limitations on data collection on platforms

In the context of first-party data collection, it is important to note that in its physical manifestation, collected data is indeed *excludable*, i.e., the data controller can restrict and control access (Schepp and Wambach, 2016[202]). Moreover, in order to obtain exclusive access, firms may implement technical measures or use contractual obligations to prevent or limit data collection by other (third-party) firms. Especially platform operators can strategically decide on how much data access they are willing to grant to independent third-parties and design their technical system accordingly. For example, media platforms that host the content of independent third-parties or marketplaces that offer products of third-party sellers can decide on the scope, the format and the granularity of data that they disclose and make accessible for third-parties.

3.1.2.4 Third-party data collection: global reach of few firms

Firms may also collect *third-party data* by tracking consumer behaviour outside of their services and applications. By doing so, firms can (i) deepen their data sets even further by collecting additional data points on their users and (ii) broaden their data sets by observing new users that have so far not adopted their services. The amount of data that can be collected along both dimensions will crucially depend on the reach of a firm, i.e., on how many services, websites and devices they can observe and trace users.

Empirical studies on the reach of third-party data collection show that although there exists a very large number of trackers, only very few firms have a significant reach. Specifically, trackers of only four firms were present on more than 10% of websites on the World Wide Web in 2016. However, the most widely encountered company, Google/Alphabet, was active on more than 70% of websites, followed by Facebook with about 30% of websites (Englehardt and Narayanan, 2016[203]). Very similar results are obtained by Ghostery, a browser extension that blocks third-party trackers.[204] The situation is likely to become even more pronounced as browsers have announced to disallow third-party cookies (see Section 2.3). This may be viewed as a step to bolster Google's and Facebook's dominance in web tracking (Financial Times, 2020[205]), because these companies have alternative means to track users across the web, e.g., through services such as 'Google Analytics' or 'Login with Facebook'.

Finally, to create meaningful user profiles from third-party data collection, a firm must be able to link and associate the collected data to individual users or specific groups of users. Here again, the probability to correctly identify users outside of a firm's property will increase with the size of its existing user base, as this allows to match identifiers or fingerprints.

---

[202] Schepp, N. P., & Wambach, A. (2016). On big data and its relevance for market power assessment. *Journal of European Competition Law & Practice*, *7*(2), 120-124.
[203] Englehardt, S., & Narayanan, A. (2016, October). Online tracking: A 1-million-site measurement and analysis. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 1388-1401).
[204] See https://www.ghostery.com/study/ and Macbeth, S. (2017). Tracking the Trackers: Analyzing the Global Tracking Landscape with GhostRank. Available at: https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf
[205] Financial Times (2020). 'Cookie apocalypse' forces profound changes in online advertising. Available at: https://www.ft.com/content/169079b2-3ba1-11ea-b84f-a62c46f39bc2?shareType=nongift

3.1.3 Data collection from external sources

Data may also be purchased from external data sources, most notably data brokers that collect and combine individual-level data from various domains (FTC, 2014[206]). These data sets usually contain the contact information of individuals (such as name and address), personal characteristics (such as age and marital status) as well as attributes of commercial interest (such as income and purchase histories). This data is frequently used to augment information about existing users, e.g., by matching users' home address and adding data about their income and creditworthiness. The data may also be used to identify audiences with specific characteristics or interests. Moreover, data brokers may provide additional services based on more fine-granular data. For example, data on individuals' credit card transactions or their order histories from e-commerce retailers are used to offer fraud detection and risk-mitigation services, which, e.g., support banks when deciding about the creditworthiness of a customer. However, in these cases, firms regularly do not obtain access to the raw data on a fine-granular level, but only receive aggregated, inferred information that is tailored to the specific purpose of the service.

Moreover, Neuman et al. (2019)[207] find that user profiles collected and created by data brokers often suffer from low quality, which limits the economic value that can be extracted. Specifically, the authors evaluate data brokers' ability to accurately infer consumers' demographic characteristics and interests based on their own data resources. They show that data on demographic characteristics is often inaccurate and may lead to wrong classifications of individuals' age and gender. For the latter attribute, data brokers' average prediction accuracy is even worse than a random guess. Accuracy is found to be higher for interest-based targeting, however, firms usually do not get access to the underlying fine-granular data, but can only access information in aggregate (e.g., in the form of "user X is interested in sports").

3.2 The breadth of data: representative data across the user base

The additional economic value that can be generated from larger data sets and the implications for competition in digital markets represents a contentious policy topic. In particular, it has been debated whether economies of scale in data processing will erect entry barriers, which shield incumbents from actual and potential competition. Next to the cost structure of the supply side, this depends on the benefits that can be achieved from more data in terms of quality improvements and prediction accuracy.

As illustrated by the three case studies in Section 2, additional data can contribute value along two basic dimensions. Firstly, data collected from different users or collected without the ability to trace users adds to the breadth of a data set (i.e., the number of columns in Figure 2 above). For example, for a search engine, observing a search query and a user's clicks on the respective search result page will provide implicit quality feedback for the observed query and the contained keywords. The more often the same query is observed, the more data becomes available for improving the search quality of this specific and possibly related search queries. Most of the policy debate and empirical studies refer to this dimension as the scale of a data set. Secondly, the same search query may also add to the depth of a data set if the search engine can trace the searcher and link the new observation to her user profile. This individual-level depth of data may contribute additional value as we will discuss in Section 3.3.

3.2.1 Positive, but diminishing returns

In the context of search engines, practitioners from both ends of the policy spectrum have suggested opposite views on the benefits of large-scale, broad data sets for the quality of search. Varian (2016)[208] points to the creation and processing of web index data as well as the development of algorithms as the key inputs for search quality. In contrast, McAfee et al. (2015)[209] highlight the quality improvements that can be achieved with respect to the ranking of search results by

[206] Federal Trade Commission (2014). Data Brokers: A Call for Transparency and Accountability. Available at https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

[207] Neumann, N., Tucker, C. E., & Whitfield, T. (2019). Frontiers: How effective is third-party consumer profiling? Evidence from field studies. *Marketing Science*, *38*(6), 918-926.

[208] The Economist (2017). Fuel of the future. data is giving rise to a new economy. Available at https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

[209] McAfee, P., J. Rao, A. Kannan, D. He, T. Qin, and T.-Y. Liu (2015). Measuring scale economics in search. Available at http://www.learconference2015.com/wp-content/uploads/2014/11/McAfee-slides.pdf.

incorporating more data on search query logs and user behaviour. They demonstrate and quantify these quality benefits from a larger breadth of search query data based on an empirical analysis of users' click-through rates on search result pages at both Bing and Google Search (He et al. 2017)[210].

They find that for search queries that have previously not been recorded by the respective search engine, click-through rates rise as more search queries are observed, which they attribute to better ranking decisions in consequence of accumulating search query data. Learning better-ranking decisions from data is the fastest for low levels of data, while quality improvements diminish as queries are observed more often. On average, they find an increase of click-through rates on the order of 2-3% over the first 1,000 queries for a new query keyword with average click-through rates for long-tail queries ranging around 70%. Similarly, Schaefer and Sapi (2019)[211] find a positive relationship between the number of searches and the quality of search results based on their analysis of Yahoo search query logs. They show that search quality improves as more users search a keyword, which resembles a direct network effect, and this quality improvement is reinforced as more individual-level data on searchers becomes available (i.e., as the depth of the available data increases; see Section 3.3).

Next to the direct quality effect for a specific search query from observing the query more often, He et al. (2017) also identify an indirect effect for improving the search quality of other search queries. In particular, a broader data set of search query data in consequence of more total observed search queries increases the probability that the query data set contains a suitable match for an entirely new or rarely observed search query. In this case, the feedback data for the existing query record can be used to improve the ranking for the new search query. Therefore, observing an additional search query does not only contribute a direct quality effect for future searches of the same query but also provides an indirect effect and positive externality for future searches of other, related queries.

In contrast to these results, Chiou and Tucker (2017)[212] do not find a significant effect on search quality, measured by users' propensity to return to the search engine after a query, when some search engines shortened their data-retention policies of search query logs in consequence of a recommendation by the Article 29 Working Party in 2008. Bing reduced the maximum storage time from 18 to 6 months, while Yahoo moved from 13 to 3 months, although the firm reversed its decision three years later, citing the need to offer "highly personalised services"[213]. However, it is important to note that the deletion of search query logs, which may be used as training data does not necessarily imply that insights from this data will also be removed from the trained algorithm used in production. The empirical results may also indicate that timeliness of data could represent a relevant criterion with respect to the economic value that can be obtained from search query logs (see Section 3.5).

Additional empirical evidence for positive, but diminishing returns from scale are reported by Bajari et al. (2019)[214] for the forecasting accuracy of future demand of products offered by Amazon with respect to the amount of available timeseries data of products. Based on internal data spanning across multiple product categories and several years, they show that collecting data on a product's sales history over a longer period increases forecasting accuracy. These benefits diminish as more data is accumulated over time, but remain positive for long timeseries of over 200 weeks. In the context of demand forecasts, it is important to consider that a longer collection period also implies that older data may become inaccurate, which could bias the findings downwards. In contrast, algorithmic improvements and additionally considered features in the data may account for some of the quality improvements, thus possibly biasing the estimates upwards.

---

[210] He, D., Kannan, A., Liu, T. Y., McAfee, R. P., Qin, T., & Rao, J. M. (2017, December). Scale Effects in Web Search. In *International Conference on Web and Internet Economics* (pp. 294-310). Springer, Cham.
[211] Schaefer, M., & Sapi, G. (2019). *Data Network Effects: The Example of Internet Search*. Working Paper. Available at https://drive.google.com/file/d/1RRxhTW560PwtMGLEN-0wHikW7oVS9CEn/view?usp=sharing
[212] Chiou, L., & Tucker, C. (2017). *Search engines and data retention: Implications for privacy and antitrust*. Working Paper. Available at https://www.nber.org/papers/w23815
[213] https://web.archive.org/web/20170224230903/http://www.ypolicyblog.com/policyblog/2011/04/15/updating-our-log-file-data-retention-policy-to-put-data-to-work-for-consumers/
[214] Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019, May). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings* (Vol. 109, pp. 33-37).

An increasing, but concave relationship between prediction accuracy and the breadth of available data set is also consistent with the theoretical predictions for several algorithms and use cases (Bajari et al. 2019, He et al., 2017). With respect to the implications for competition, this calls for analyses of whether current applications in practice are still subject to significant quality improvements when increasing the breadth of the underlying data set and, in reverse, how quickly service quality reaches a level where benefits from additional data are only marginal.

3.2.2 The sparsity of observed data on user behaviour

In this context, several simulation and empirical studies emphasize that the *sparsity* of a data set has important implications for the improvements in prediction accuracy that can be achieved by augmenting the scale of a data set through additional observations. In general, a data set is considered sparse if, for any user, the vast majority of data entries (i.e., the rows in Figure 2) are empty, because they have not been observed before. Observed data on fine-grained user behaviour is regularly sparse as the data "consists of individually rare, but collectively frequent events" (Halevy et al., 2009[215], p.9). For example, at an online retailer, most products will only be purchased from a small number of customers, whereas the majority of customers will not interact with these products.

Li et al. 2016[216] show that higher sparsity will lead to a lower convergence rate of prediction performance, i.e., diminishing returns to scale will set in only for higher levels of data breadth. Moreover, both Junqué de Fortuny et al. (2013)[217] and Martens et al. (2016)[218] demonstrate that prediction accuracy increases for larger (broader) data sets of fine-grained, behavioural user data. Whereas benefits decrease marginally as prediction accuracy approaches the theoretical benchmark, the studies show that this convergence has not been reached in many popular application settings at that time.

Amatriain (2013)[219] highlights that with respect to the value of data for prediction tasks, it is important to distinguish between models with high variance and models with high bias. A high variance occurs especially for complex models that include a lot of features and can be addressed by more training data or by reducing the number of features of a model. Specifically, in these cases "The Unreasonable Effectiveness of Data"[220] may generate large-quality improvements. In contrast, simpler models with a smaller number of features will not equally benefit from more data as the maximum performance is reached more quickly for a smaller (thinner) data set.

Whether more features will increase prediction quality depends on the specific task, although the recently introduced machine learning techniques in many domains generally rely on larger feature sets than past algorithmic approaches. This holds especially for search and recommendation systems that take into account data across services and domains. Even in cases where firms may need to reduce the size of their input data sets due to computational constraints, sampling or approximation methods that reduce data volume, but retain relevant properties of the original data set, benefit from a larger input data set (see, e.g., Wedel and Kannan, 2016[221]).

3.2.3 Implications

Concerning the analysis of data, empirical studies suggest that there are benefits from more data especially with respect to prediction accuracy in many use cases, but these benefits are marginally decreasing as the scale and breadth of data sets increases (diminishing returns to scale). Whether learning from additional data will still provide significant quality gains or whether diminishing returns are reached rather quickly is influenced by the sparsity of data and the complexity of the employed algorithms. For sparse data sets and complex prediction algorithms with many features, there is evidence that learning from additional data remains significant over larger data sets. Increasing

---

[215] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.
[216] Li, X., Ling, C. X., & Wang, H. (2016). The convergence behaviour of naive Bayes on large sparse datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(1), 1-24.
[217] Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: is bigger really better? Big Data, 1(4), 215-226.
[218] Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining Massive Fine-Grained Behaviour Data to Improve Predictive Analytics. MIS Quarterly, 40(4), 869-888.
[219] Amatriain, X. (2013). Mining large streams of user data for personalised recommendations. *ACM SIGKDD Explorations Newsletter*, *14*(2), 37-48.
[220] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.
[221] Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97-121.

returns to scale in the early learning period will furthermore play a relevant role for prediction tasks where previously observed data outdate quickly.

The case of search engines also highlights that diminishing returns from scale and quicker learning for queries with few data have more differentiated implications, as scale can also have indirect benefits for rare queries with few data (He et al, 2017). On the one hand, positive, but diminishing returns from observing more queries could imply that moderate scale is sufficient to achieve most of the potential quality increases. On the other hand, it implies that for long-tail queries that are observed rarely, the number of similar queries from which a corresponding match may be identified matters even more. There is thus an indirect benefit of breadth, as this increases the probability of finding an appropriate match in other observations, which may be incorporated to improve quality.

In this context, it is important to note that diminishing returns to scale for a given prediction task may not rule out that there exist positive feedback loops and externalities (see Schaefer and Sapi, 2019) that may establish a competitive advantage for incumbent firms with large data sets and raise entry barriers (see Section 3.4).

Finally, the empirical results reviewed in this section indicate that broader data collection gradually transforms into quality improvements, albeit at a decreasing marginal rate, which is in line with the observations from our case studies. Only in specific cases, there seems to exist a *minimum viable scale* with respect to the amount of data that is required. For the online advertising industry, for example, Lewis and Rao (2015)[222] find that only very large amounts of data allow firms to measure whether advertising campaigns are indeed successful.

3.3 Depth of data: detailed data on individual users

If new information is collected as pseudonymous or personally identifiable data, this adds to the depth of a data set. In both these cases, new information can be linked to the existing data about an individual user and hence user profiles can be created or enriched. In contrast, anonymous data, by definition, cannot be traced back to an individual and thus cannot be linked to previous observations and volunteered information. Nonetheless, anonymous data collection still adds to the breadth of data.

Whereas policy debates and privacy legislation have mostly centred around issues concerning personally identifiable data, what matters arguably more in most use cases from a technical and economic perspective is whether data is traceable, which is already the case for pseudonymous data. This is because predictive performance and economic benefits usually rely on the linked information contained in a user profile and not on the personal identity behind that profile. Sufficiently deep and granular data will often provide the same or even more accurate information than direct personal identifiers such as the name or address of an individual. Of course, it is often necessary to ultimately re-identify an individual user, e.g., in the case of targeted advertising. But even in these cases, it is sufficient to sync identifiers rather than to get to know the individual personal identity. In any case, the depth of data does play an important role with respect to the economic value of the entire data set as demonstrated by recent empirical studies.

Based on data from the search engine Yandex, Yoganarasimhan (2020)[223] shows that the quality of search results due to personalisation increases the more data is available on an individual user's search history. In the specific use case, personalisation improves click-through rates on the top position of the search results significantly by 3.5 percentage points on average. Quality improvements are more substantial for long-term personalisation across search session than for short-term personalisation within a session, indicating that the long-term traceability of users is important for achieving quality improvements. The study finds that search quality metrics gradually improve with growing lengths of user profiles and relative benefits are especially large when data becomes available for users with short observed search history. Although returns from more data decrease with deeper user profiles, they are still growing for the largest observed user histories in the data sample.

[222] Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics, 130(4), 1941-1973.
[223] Yoganarasimhan, H. (2020). Search personalisation using machine learning. *Management Science*, *66*(3), 1045-1070.

Overall, ranking signals based on data of individual user profiles account for more than 50% of the improvement from personalisation and about 28% of quality improvements are generated by data on users' implicit click-based feedback (Yoganarasimhan, 2020). Moreover, it is shown that benefits of personalisation are also influenced by the availability of data on general user behaviour, as website- and domain-specific data contribute more than 20% of quality improvements. It is also found that the depth of user-profiles is especially beneficial for search queries with a higher variety in users' clicks on search results. Therefore, individual-level data is more valuable for transactional and informational search queries, where the intent of different users are likely to vary depending on the context of their search, than for navigational queries, where most users agree on the ranking order of search results. More generally, this highlights that increasing the depth of data can generate quality improvements that are not achievable by simply increasing the breadth of data.

In an online field experiment, Claussen et al. (2019)[224] test whether a personalised recommendation system at an online news website can outperform human editors with regard to users' likelihood to engage with suggested news articles. They find that clicks to articles recommended by the algorithm increase with the number of prior visits of users, i.e., with the amount of data that becomes available about the observed behaviour of the user. After about six to ten visits, the recommendation system outperforms the human editor in terms of increasing user engagement indicating that the collection of individual-level observed data generates economic benefits. Improvements in recommendation accuracy diminish with more individual-level data and level off at about 50 visits of a user. The authors emphasize that their findings suggest continuous quality improvements without significant threshold effects or minimum viable scale effects. Due to the experimental setup, the analysis by Claussen et al. (2019) focuses on the direct effect of more individual-level data, i.e., increasing the length of an individual user profile. Thus, they do not consider the indirect effects of additional observations on recommendations to other users.

Based on data from the search engine Yahoo, Schaefer and Sapi (2019) highlight that increasing the depth of data may have an important indirect effect on the benefits from increasing the breadth of data. More specifically, they show that the ability to track user identities across search sessions, e.g., by the means of browser cookies, accelerates the speed at which a search engine can improve the quality of its search results when observing additional search queries. In particular, a longer sequence of observed search sessions from an individual user leads to better ranking decisions as measured by the users' click-through rate. But a deeper data set does not only offer a better prediction for the individual user herself, it also facilitates the identification of similar user profiles, which further improves prediction accuracy for other future search queries. Hence, there is a positive externality from adding a data point on a specific user, for the future requests of all users. A deeper data set means that similar users can be identified more precisely and hence more accurately. Thus, depth may contribute additional economic value even if there are already diminishing returns for increasing the breadth of data. These indirect effects will be especially important for new search queries, for which no (broad) item-specific data is available, as quicker learning is particularly important for these queries.

Taken together, the empirical findings suggest that increasing the depth of data contributes additional value beyond the benefits that can be achieved from increasing the breadth of data. Along the depth dimension, empirical findings suggest that there are continuous, but diminishing returns from additional data, although the convergence of these benefits may vary significantly depending on the specific use case and context. Most notably, there is also an indirect effect of depth, such that additional data about an individual user also contributes to the economic value creation from increasing the breadth of data. This implies that data sets on user behaviour, which contain (pseudonymized) identifiers, can provide significant additional value over anonymised data sets. Moreover, individual traceability can contribute significant economic value even when the information does not explicitly identify an individual's characteristics. Finally, the reinforcement effect between both scale dimensions may also rationalize data-driven feedback effects from data even in the absence of increasing returns to scale along both the breadth and depth dimension.

[224] Claussen, J., Peukert, C., & Sen, A. (2019). The Editor vs. the Algorithm: Targeting, Data and Externalities in Online News. Working Paper. Available at https://dx.doi.org/10.2139/ssrn.3399947.

3.4 Data-driven network effects

Beyond direct and indirect network effects from the number of users, it has been suggested that data-driven network effects can give rise to self-reinforcing feedback effects from a dynamic perspective. Specifically, two conceptual feedback loops are frequently conjectured (Lerner, 2014[225]; Bourreau et al., 2017[226]): Firstly, the more consumers are using a service, the more (volunteered and observed) data is created on which data analytics can be performed and algorithms can be trained, which in turn results in an improvement of the service (e.g., more accurate recommendations and more relevant search results), which in turn leads to more consumers (user feedback loop). Argenton and Prüfer (2012)[227] argue that such feedback loops emerge with regard to search engine's collection and analysis of query logs, which serve as a main input for the improving the ranking of search results as described in Section 2.1.

Secondly, more data in consequence of more users also enables more effective targeted advertising and thus generates larger advertising revenues. In turn, this allows the firm to invest more in service quality or other added value for consumers. Again, this is assumed to result in more users (monetization feedback loop). In digital markets, the indirect generation of revenues on other market sides often allows platforms to set prices to zero on the consumer side. Thus, market entry and reaching a critical scale may require large and long-term investments from new market participants if the incumbent benefits from a monetization feedback loop.

Given these feedback effects, economic benefits from data translate into competitive advantages. In particular, first mover advantages can become sustained competitive advantages, because competitors are unable to initiate the same feedback loop. Specifically, the lack of access to data that is created by fellow users – a type of indirect network effect – creates a barrier to entry. In consequence, the lack of data limits the ability of a new service to compete on the basis of algorithmic insights and data analytics. This argument is explored more formally, for example, in Hagiu and Wright (2020)[228], who show that this competitive advantage of the incumbent prevails under various assumptions about the shape of the learning curve from data.

With regard to the empirical findings reported in Sections 3.3 and 3.4, it is important to recognize that diminishing returns to data scale in either the breadth or depth dimension do not automatically imply that data-driven network effects cannot create substantial entry barriers. The empirical results indicate a reinforcing effect between the two dimensions (Schaefer and Sapi, 2019). As learning from additional search queries of different users (breadth) is accelerated when more individual-level information is available on these users (depth), positive feedback loops in consequence of data-driven network effects can emerge even in the absence of increasing returns to scale. As we will discuss in Section 4.1, such data-driven network effects can have significant ramifications for the competition within and across digital markets.

3.5 Quality of data

Generally, the *quality of data* can be measured along the following dimensions:

- Accuracy, referring to whether the data correctly represents the facts.
- Timeliness, referring to how fast can data be collected and how quickly it becomes outdated.
- Granularity, referring to the level of detail of the data collected on individual interactions.

3.5.1 Accuracy

Accurate data is necessary to infer correct information and knowledge from data. Hence, it is sometimes argued that accuracy is more important than the size of a data set (Schepp and Wambach, 2015). High inaccuracy can render even large-scale data sets worthless. However, it is also a characteristic of "big data" that the increasing volume of data usually facilitates to identify inaccurate data points and to obtain robust predictions even with lower data accuracy.

[225] Lerner, A. V. (2014), The Role of 'Big Data' in Online Platform Competition (August 26, 2014). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2482780
[226] Bourreau, M., A. de Streel, and I. Graef (2017). Big Data and Competition Policy: Mar- ket power, personalised pricing and advertising. Available at http://www.cerre.eu/ publications/big-data-and-competition-policy
[227] Argenton, C., & Prüfer, J. (2012). Search engine competition with network externalities. *Journal of Competition Law and Economics*, *8*(1), 73-105.
[228] Hagiu, A. and Wright, J. (2020). Data-enabled learning, network effects and competitive advantage. Mimeo. Available at: http://andreihagiu.com/wp-content/uploads/2020/02/Data-enabled-learning-20200214_web.pdf

With respect to data collected from users, it is important to recognise that the accuracy of data may inherently differ between the modes of collection, i.e., between volunteered data and observed data. As volunteered data is derived from direct human input, this data may more often be inaccurate, e.g., because wrong information (such as a wrong email address, fake name or fake review) is submitted intentionally or unintentionally. However, as described in Section 3.1.2.2, in cases where volunteered data is collected as part of a service that requires accurate information for fulfilment or effectiveness of the service (such as in the case of e-commerce or personalised recommendations), users will have a strong incentive to provide accurate information. In contrast, consumers may abstain from adopting a new service if this requires the disclosure of personal data, but there is insufficient trust in the new service provider among consumers.

By contrast, observed data is less prone to deliberate manipulation, because it is derived from actual behaviour and sensors. The accuracy of the observed data is still very context-dependent. For example, click data from an e-commerce session can be very noisy and sparse because the user might just be browsing through random products and in each product category only very products are explored. In another session, the similar click data can be very accurate and dense, as a consumer explores several similar products and puts some of them in the shopping basket, but finally only buys one. Similarly, data from sensors (e.g., GPS sensors) can be highly accurate at times and inaccurate at other times.

In cases where digital services aim to predict users' preferences or require inference of users' actual intent, observed data is often more accurate than volunteered data. This is because observed data can reveal implicit information and context-specific peculiarities that users may not even be aware of. Moreover, in online search and e-commerce, consumers are often uncertain about their actual objective. In consequence, volunteered information may often be inaccurate, whereas sufficient behaviour data can facilitate inference of the true intent.

3.5.2 Timeliness of data

Concerning the data inputs of today's digital businesses, it has been emphasized that the access to data flows rather than to data stocks is relevant (Davenport et al., 2012[229]). Because the economic value from data often depends on accurate predictions of consumers' preferences, which may evolve over time and with varying product and service offers, timeliness of data represents an important quality dimension. Moreover, economic benefits may directly be tied to the knowledge of a consumer's current situation. For example, in the case of behavioural targeting or local search queries, precise data on the current context of a user are required to provide adequate matches that satisfy the user's current intent or information need. Hence, the value of data may often be transitory and depreciate quickly for several applications (Sokol and Comerford, 2017[230]).

This has led to conclusions by some observers that competitive advantages from data are rather limited in these cases, because the underlying data may become outdated quite quickly (see, e.g., Schepp and Wambach, 2015). Moreover, older data could become inaccurate as the information changes over time and thus large data sets that are collected over a long period may more likely be of lower quality. Thus, according to this line of argument, competitors do not need to duplicate the entire data stock of an incumbent but can focus on collecting only the most relevant (current) data. Moreover, the incumbent is under pressure to innovate and continue data collection rather than being able to exploit and monetize its existing data stock (Krämer and Wohlfarth, 2018[231]).

However, the need to continually update data may indeed reinforce the advantages of firms that are in a superior position to observe consumer behaviour. If timely data is required as in the case of local search or behavioural advertising, competitors will find it more difficult to collect data from alternative sources. In particular, advantages of firms that can collect observed (behavioural) data rather than having to rely on volunteered data, e.g., in the form of explicit feedback by consumers, are fortified if the value of data outdate quickly. In turn, this further limits the extent of non-rivalry

[229] Davenport, T. H., Barth, P., & Bean, R. (2012). How Big Data Is Different. *MIT Sloan Management Review*, *54*(1), 22-24.
[230] Sokol, D. D., & Comerford, R. E. (2017). Does antitrust have a role to play in regulating big data? *Cambridge Handbook of Antitrust, Intellectual Property and High Tech*. Cambridge University Press.
[231] Krämer, J., & Wohlfarth, M. (2018). Market power, regulatory convergence, and the role of data in digital markets. *Telecommunications Policy*, *42*(2), 154-171.

in data collection, making it harder for other firms to duplicate data assets, as a consumer will regularly allow only a limited set of devices and services to observe his actions continuously.

In use cases where services must react instantly to a user inquiry, the value from data is highly transitory. But precisely in these cases the benefits of being able to collect data about an individual user, possibly from multiple sources, are especially pronounced. For example, knowing about a user's current location or her current mode of movement drastically increases prediction performance with respect to personalised recommendations as well as search queries, because the context of a user request can be inferred more accurately. In these situations, it is very difficult to collect alternative data that could provide a similar quality with respect to the requirements of instant feedback tasks. Next to individual-level data, this also applies to the collection of anonymous or pseudonymous data. Thus, in cases where data is transitory, the ability to collect broad data sets is especially valuable. For example, aggregated live feedback from mobile devices may provide overall traffic information, but the accuracy of such information depends crucially on the number of devices that can be observed.

With respect to the relevance of data flows, data quality is thus affected by the resolution at which data can be collected. Not only does a higher resolution guarantee fresher data, but it may also allow for the recognition of more nuanced patterns in the data. For instance, repeating user behaviour may inform product recommendations, but also forecasts about future demand in e-commerce applications.

3.5.3 Granularity

Data quality can also be measured with respect to the granularity of the data, which is determined by the level of detail of the collected data. Figure 3 illustrates varying levels of granularity for the example of search query log data. In this vein, one can think of granularity as of how many individual data points are collected and stored to record an interaction with a service. For example, a search query log may only store the keywords of the query itself (case a), or additionally capture the URL of the search result that the user clicked on (case b). To link a search query to additional context data the search query log may also record a session identifier (case c). Finally, the location of a user may be collected to add further detail on the context of a respective search query.

The added economic value from a higher granularity strongly depends on the specific context. Highly accurate GPS data, for example, may be necessary to identify which products a consumer was interested in when visiting a department store, whereas coarser data may still be acceptable to identify which stores a consumer has visited in a mall. However, the case studies in Section 2 have illustrated that more fine-granular data usually allows for better prediction performance and thus larger economic benefits, if indeed the available data is sufficiently broad and deep. As we will discuss in Section 5.2.4, granularity is a key dimension to balance the trade-off between competition goals and privacy issues when devising data access remedies.

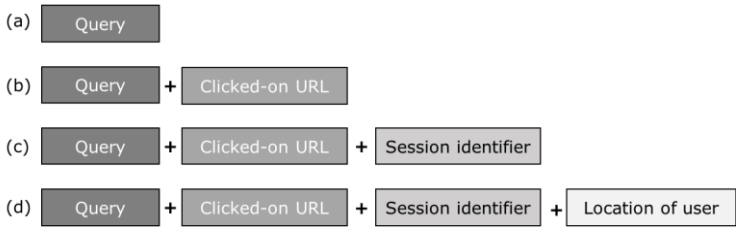

Figure 3: Varying levels of granularity (level of detail) of search query log data.

3.6 Complementary data assets

The collection and processing of data requires complementary input resources. Firms must deploy and operate the necessary information technology (IT) infrastructure to support the storage and analysis of large data sets. This includes distributed databases and physical servers, which are usually deployed in data centres across the geographic regions of operations. The case studies in Section 2 highlight that in many applications, services require immediate access to the stored data (e.g., online search engines must match user queries against the search index in milliseconds), which, next to the computing facilities, necessitates network infrastructure that ensures reliable connectivity with high throughput and low latency. Thus, data-intensive applications regularly require firms to undertake considerable investments into the physical infrastructure as well as virtual technologies, such as data structures, that support the storage, distribution and processing of data.

Empirical studies suggest that such IT infrastructure investments have indeed a positive impact on firm performance. Investments into big data and analytics assets (e.g., in the form of database technologies analytics software) have been found to significantly improve firm productivity and financial performance (Müller et al., 2018[232]). Moreover, empirical evidence suggests that productivity gains are driven by a complementary relationship between IT infrastructure and data (Brynjolfsson & McElheran, 2016a[233]). This confirms the intuition of a reinforcing relationship between IT infrastructure and data inputs. On the one hand, firms with access to valuable data resources have a higher demand for the necessary IT capabilities and are thus incentivised to develop and expand their IT infrastructure, while, on the other hand, firms with a well-developed IT infrastructure are more likely to exploit the data resources at their hands.

Due to the high fixed costs of IT infrastructure assets, associated *economies of scale* (see, e.g., Brynjolfsson & McElheran, 2016b[234]) can extend competitive advantages of data-rich firms and establish entry barriers for market entrants. Moreover, complementarities between data and technical infrastructure give rise to significant supply-side *economies of scope*. For example, Amazon has become a leading provider of cloud and web services based on its IT engineering continuously expertise and its distributed computing infrastructure, originally developed to support its global e-commerce platform. In consequence of scope advantages, data-rich firms may not only benefit from their access to data when entering adjacent or new markets, but also from their established IT infrastructure assets.

Scope and scale advantages of large firms with regard to IT infrastructure assets may be mitigated by the availability of cloud-based IT services. Such services can reduce the initial fixed costs that a firm must incur to set up the necessary IT infrastructure and facilitate the dynamic scaling of this infrastructure according to the firm's need. Thus, cloud services can facilitate market entry by reducing the scale and scope advantages of large incumbents. Nonetheless, the usage-based pricing for such on-demand IT infrastructure services will still put the entrant at a continued cost advantage vis-à-vis the incumbent unless it invests in its infrastructure assets.

Next to hardware and software infrastructure inputs, data must be analysed and processed to infer actual knowledge and create economic value (see the case studies in Section 2 for applications across

---

[232] Müller, O., Fay, M., & vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, *35*(2), 488-509.
[233] Brynjolfsson, E., & McElheran, K. (2016). Data in action: data-driven decision making in US manufacturing. Available at https://dx.doi.org/10.2139/ssrn.2722502
[234] Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, *106*(5), 133-39.

different domains). Thus, algorithms and skilled human resources represent important complementary inputs to extract economic value from data besides IT infrastructure assets. In this context, authors have argued that algorithms and the technical know-how of employees and organisations represent the key factors to sustained competitive advantages rather than the access to the actual data resources (Lambrecht & Tucker, 2013). Besides, it has been suggested that entrants may challenge data-rich incumbents based on innovative algorithms that rely on smaller, but more specialised data sets of higher quality (Schepp & Wambach, 2016).

It has been found that firms, which invest in employees with data skills, exhibit faster productivity growth, however, only if these firms also have access to significant data assets (Tambe, 2014). This points to a complementary and reinforcing relationship between data resources and data-skilled employees (see also Brynjolfsson & McElheran, 2016a). Moreover, whereas algorithmic approaches based on smaller or alternative data sets may in some cases serve as substitutes to algorithms relying on larger data sets, combining algorithmic innovations with larger data sets will in general lead to superior overall performance (see, e.g., Martens et al., 2016). This is especially the case for machine learning approaches as discussed in the individual case studies and also in Section 3.2 (see also Halevy et al., 2009). Hence, due to the complementary relationship of inputs, superior algorithmic design on its own is unlikely to compensate for a lack of access to data in the long run.

Instead, the development of innovative algorithmic approaches will often depend on having access to large-scale data sets. For instance, the Netflix prize, which gave independent computer scientists access to a large real-world set of user data, has triggered significant innovation with respect to recommendation algorithms (Amatriain & Basilico, 2015). In this spirit, data-rich firms like Amazon regularly attract highly-skilled researchers based on the proposition to grant access to large sets of user data (see, e.g., the Alexa Prize[235]). Hence, data on its own may indeed be of little economic value without the proper algorithmic tools, but the access to data itself can reinforce a firms' ability to attract the human resources necessary to develop these algorithms.

[235] https://developer.amazon.com/alexaprize

04

DATA-DRIVEN THEORY OF HARM AND POLICY OBJECTIVES

4 Data-driven theory of harm and policy objectives

4.1 Data-driven theory of harm

There has been much debate of the various theories of harm in the context of digital markets and the appropriate welfare standard to consider. Theories of harm may also involve a **'citizen perspective'** that is not confined to a **narrow economic rationale**, but rather based on notions of 'fairness' and 'democracy'. However, such considerations are outside of the scope of this report, whose purpose is to focus on *theories of harm that relate to data*, and on the *economic regulation of digital markets.* In the following, we focus on a discussion of the theory of harm related to data more precisely but leave aside other possible theories of harm for which we refer to excellent existing surveys (see, e.g., Cremer et al, 2019, Scott Morton et al , 2019 and Parker, Petropoulos and Alstyne, 2020[236]).

In principle, three arguments are being made here in relation to data-driven theories of harm.

- Firstly, in cases where data-driven network effects (see Section 3.4) are strong, markets tend to monopolize (**market tipping**). In the process, the monopolist amasses more data, wealth and skilled labour, and thereby constitutes effective **entry barriers**.
- Secondly, this tipping effect does not stop in the very market where it started but may **spill over to related, data-intensive markets**, which can already exist or may still emerge. As more and more user data is collected in more and more markets, the data-driven network effects are likely to become stronger and stronger.
- Thirdly, this also affects **innovation**, because high entry barriers stifle innovation activity in those areas and markets where entrants may set out to compete with the incumbent. High entry barriers and a lack of competition also diminish the innovation incentives of the incumbent. Besides, there is growing evidence that these high entry barriers also lead to lack of access to venture capital, which raises entry barriers even more for de-novo market entrants. Finally, we suggest what can be realistic policy goals in view of these arguments.

4.1.1 Market tipping due to data-driven network effects/economies of scope, and the impact on innovation

In the case studies, we observed a common theme. While consumers are using the service (search, e-commerce, media), they **leave a digital footprint** by revealing personal preferences, relevance assessments of rankings, or individual behavioural biases, just to name a few. Often the service providers even design the services in such a way that keeps users engaged, and thus they contribute even more data. This *user data* can then be used to improve the service directly, e.g., by improving the quality of the search results, or the quality of the content and product catalogue, or indirectly, e.g. by improving the quality of the product and content recommendations. This is what we referred to as data-driven network effects in Section 3.4.

On a more abstract level, such **data-driven network effects (or data-driven economies of scope) allow lowering the cost of innovation**. This mechanism has been made explicit, for example, in Prüfer and Schottmüller (2017)[237], who formalize the idea that more data reduces the marginal cost of innovation, and that in consequence, an initial small market advantage can unravel, such that the market "tips" towards the firm who had that advantage. For example, a company that has access to the location data of many customers finds its easier to develop a digital maps service which also shows road traffic and busy hours of stores, which then improves of other maps services and leads to more demand, which in turn yields more location data. Likewise, a company with many users can test different designs and features with higher validity using field experiments (A/B tests), and can better infer new relevant features from user behaviour. This will, in turn, drive more demand and allow for even better analyses and forecasts. While initially, the incentive to innovate is high because high gains are expected from a tipped market, the incentive to innovate in a tipped market are low, everything else being equal. This is because entry barriers in such a market are high. The incumbent firm has both high service quality and a lower marginal cost of innovation. Thus, it can

[236] Parker, G. and Petropoulos, G. and Van Alstyne, M W., Digital Platforms and Antitrust (May 22, 2020). Available at SSRN: https://ssrn.com/abstract=3608397.
[237] Prufer, J., & Schottmüller, C. (2017). Competing with big data. *Tilburg Law School Research Paper*, (06).

easily overcome the innovation efforts of the entrant.[238] This lack of innovative pressure from entrants means that incumbent firms are themselves less likely to invest in innovation (Segal and Whinston, 2007[239]; Prüfer and Schottmüller, 2017).

This is also in line with more traditional theories of innovation, pioneered by Arrow and Schumpeter. Under an Arrowian view (Arrow 1962)[240] firms facing strong competition have a higher incentive to innovate to escape from the competition. Under a Schumpeterian view (Schumpeter 1934)[241] firms in a monopoly position have a higher incentive to innovate to protect their monopoly and to discourage entry (what Schumpeter calls 'creative destruction'). Taking both arguments together, and in line with ample empirical evidence, innovation incentives tend to be the highest in oligopolies with several firms (see, e.g., Aghion et al. 2005)[242], but not in monopolies. Consequently, in **digital markets that have tipped and where entry barriers are therefore high, the threat of 'creative destruction' is significantly reduced**. This means that monopolistic incumbents have less incentive to invest in innovation in order to protect their monopoly position, everything else being equal.

Data-driven network effects cannot be easily copied by competitors, because even if a competitor were to acquire all necessary infrastructure and skills (see Section 3.6), it would still lack the continuous inflow of user data that would be necessary to develop the same data-driven insights and to train algorithms in a way that would deliver a better product or service. For example, even in the extreme case where a potential competitor would replicate Google's massive technical infrastructure of data centres, and hire all of Google's skilled labour to replicate its (untrained) search algorithm, it would still offer a worse search engine, because it would lack the user data to train and tweak the algorithm in such a way that it would be competitive. This is especially problematic from an innovation perspective, and welfare perspective, because that would even be true if the competitor had come up with an arguably better algorithm (given it were trained with sufficient data). This data-driven network effects also discourage innovation efforts by entrants and is less likely to receive funding by venture capital (see Section 4.1.5).

4.1.2 Data-driven envelopment or "the domino effect"

Data-driven network effects also give rise to a related phenomenon, which is known as envelopment or "the domino effect". Due to data-driven network effects, **data-rich firms may also venture into other data-driven markets more easily**. On the one hand, they enjoy scale and scope effects because the infrastructure and skills necessary to store and analyse large amounts of data can be readily used in related markets, where data is also key. Although typically some domain knowledge is necessary to make sense of the data or to understand the business, a firm's superior ability in handling and analysing data, often combined with its superior insights about customer preferences, allow it to innovate better than a less (user) data-rich firm. For example, the data skills and infrastructure that Amazon has acquired to support its e-commerce activity allowed it, among other things, to venture into the streaming video market, where it could make use of its data centres, and its ability to recommend suitable content. By venturing into a related market, data-rich firms acquire access to even more users and even more data, which strengthens their data-driven network effect, and allows them to potentially venture into even more markets. This is the "domino effect" highlighted by Prüfer and Schottmüller (2017).

On a related note, Eisenmann et al (2011) [243] argue that such network effects (not limited to data-driven network effects) allow firms even to venture into other markets that also exhibit network effects, and would therefore otherwise be shielded from the competition as well. This is a process that they call "**envelopment**". A competitor that already has a large installed base of users can

[238] Here we focus on market-driven network effects, but market tipping may not just occur as a consequence of data-driven network effects, but also occurs in the presence of strong positive network effects more generally, e.g., for example due to direct network effects in a social network.
[239] Segal, I., & Whinston, M. D. (2007). Antitrust in innovative industries. American Economic Review, 97(5), 1703-1730.
[240] Arrow, K. J. (1962). Economic Welfare and the Allocation of Resources for Invention. The Rate and Direction of Inventive Activity, 609–26. https://doi.org/10.1515/9781400879762-024.
[241] Schumpeter, J. A. (1932). The Theory of Economic Development: an Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Harvard University Press.
[242] Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. The Quarterly Journal of Economics, 120(2), 701-728.
[243] Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal*, *32*(12), 1270-1285.

overcome entry barriers due to network effects by bundling its existing service with the new service. Bundling is possible even for functionally unrelated services, such as e-commerce and video streaming, as long as the two services cater roughly to the same user group. For example, in case of Amazon bundling was done through a "Prime" membership, which bundled free shipping with free access to (a part of) its video streaming service, and to which more services are added. Building on these insights, Condorelli and Padilla (2020)[244] highlight that envelopment can also occur through the tying of privacy policies. That is, by consenting to their data use in one service, consumers also have to consent to their data use in another, possibly unrelated, service of the same firm. This facilitates data-driven network effects and the domino-effect.[245]

In a world where more and more businesses undergo digital transformation, more and more markets become prone to the domino effect or envelopment. In the past decade or so, tech giants have mainly confined to venturing into the digital sphere, but there is reason to believe that the next wave of envelopment will venture **more into the physical sphere**. Areas that immediately come to mind are health, farming, energy, logistics or autonomous driving, just to name a few. In this context, data-rich incumbents in the digital sphere have recently expanded their data collection in physical spheres as highlighted in the case studies concerning local search and recommendations by voice assistants on various home automation devices.

Thus, under these favourable conditions and without further safeguards, the domino effect and envelopment is likely to continue. Following this logic, more and more markets, both digital and those that undergo digital transformation, are likely to become more concentrated, and therefore possibly less competitive and innovative. We do not suspect, however, that this process can continue indefinitely, as next to data-driven network effects, which drives conglomeration, **firm size is also limited by other constraints, such as management issues and other 'transaction costs'** of doing business.[246]

It is worth highlighting that in the preceding discussion we have assumed that this occurs even without any evident misconduct, such as the abuse of a dominant position through self-preferencing. Such practices may expedite the process and provide an additional barrier to entry for competitors. As many commentators have noted, **competition law has proven to be too slow for effective interventions in this very dynamic process**, because it is virtually impossible to find remedies that can restore the data-driven network effects to what they were before the abuse. These arguments reinforce the reasons for ex-ante regulation that complements ex-post competition law.

We should also note that we have focused on data-driven effects here. More generally, in digital markets, one also often finds **traditional network effects**. These can be either direct, such as in social networks or messaging services, or indirect, such as in two-sided markets (e.g., Amazon marketplace). Moreover, digital firms tend to create economies of scope and scale for consumers through building an interoperable ecosystem of services and devices. For example, Apple's smart watch can only be used with its iPhone, or Amazon's FireTV or Tablet has Alexa and Amazon Prime Video readily installed. These practices add to the envelopment and concentration trends; and they raise entry barriers further which likely harms innovation (see, e.g., Choi and Stefanadis, 2001[247]). However, this is not in the focus of this report and for a more complete treatment of such 'conglomerate effects' we refer to Bourreau and de Streel (2019)[248].

4.1.3 Envelopment revisited: Ancillary data services

The case studies revealed also a more subtle means of data agglomeration exercised by several data-intensive firms. Instead of venturing into related markets themselves to gain more user data to fuel data-driven network effects, **often ancillary services are offered to third-parties, which can act as a conduit for data exchange**. For example, recently identity management services like "Login with Facebook", "Login with Google" or "Sign in with Microsoft", to name only a few, have

[244] Condorelli, D., & Padilla, J. (2020). Harnessing Platform Envelopment in the Digital World. Journal of Competition Law & Economics.
[245] We discuss this in more detail in Section 5.1.1.
[246] Holmstrom, B., & Roberts, J. (1998). The boundaries of the firm revisited. *Journal of Economic Perspectives*, *12*(4), 73-94
[247] Choi, J. P., & Stefanadis, C. (2001). Tying, investment, and the dynamic leverage theory. *RAND Journal of Economics*, 52-71.
[248] Bourreau, M., & De Streel, A. (2019). Digital conglomerates and EU competition policy. *Available at SSRN 3350512*.

become popular. These are offered now by nearly all major digital platforms.[249] In a similar vein, more and more platforms are offering payment services to third-parties, such as "Pay with Amazon", "Pay with Paypal" or "Pay with Apple". From a data perspective, the common theme in all of these ancillary services is that they offer third-parties a distinct benefit, often to the extent that the third-parties even receive some (accurate) user information at first, such as their address, name or email. However, by using these ancillary services, the third-parties may also reveal user data to the providers of the service, namely about when and to which service they logged on or which products or services they bought (see, e.g., Preibusch et al., 2015[250]). For the data-rich platform provider, the user data that it gets in return is often far more valuable than the (loss of exclusivity to the) data that it has revealed. In this way, users can be tracked and traced even on those websites and services that are not operated by the service provider itself. Similar reasoning applies for tracking technology like "Google Analytics" or "Facebook Pixel", which helps third-party websites to gather more insights about their audiences, but also allows the provider of these services to gather more user data from across the web, and at even greater detail than the data that is disclosed to the third-parties. Indeed, already four years ago, Englehardt and Narayanan (2016) found that Alphabet/Google, and to some extent Facebook, had access to tracking data from an extensive reach of websites, which could not be matched by any other firm (see Section 3.1.2.4).

Such data collection spurs data-driven network effects and equally contributes to raising entry barriers in data-intensive markets. More specifically, the theory of harm associated with the practice of offering ancillary services to collect more data is formalised by Krämer, Schnurr and Wohlfarth (2019)[251]. They highlight that independent websites can end up in a dilemma, whereby competitive pressure forces them to adopt such ancillary services to gain a competitive advantage over other independent websites with whom they compete immediately for users in a given market. However, as other websites adopt the ancillary service as well, eventually no website can gain a true competitive advantage. Instead, the adoption of the service has resulted in a permanent data transfer to the service provider, which results in less exclusivity of user data and a weakened competitive position in the broader market for users' attention. The contentious issue about this practice is that, **in the short run, adopting the ancillary service is often welfare increasing, both from a consumer welfare and a total welfare perspective. However, the independent websites are worse off than if they had not adopted the service**, hence creating a dilemma for them. Thus, from a dynamic welfare perspective, competition in the data-driven economy is weakened, and the competitive position of the data-rich service provider is strengthened further.

4.1.4 Vertical integration and data use

Similar issues as in the case of ancillary services arise in case the **data-rich incumbent is vertically integrated and offers its downstream service on the same platform**. This is the case in all three case-studies that we reviewed: in e-commerce marketplaces, e.g., because Amazon operates both the platform and acts as a seller on the platform; in search, e.g., because Google operates both the search engine and operates other services that are findable through the search engine, such as a shopping comparison service or a maps service; in media, Google operates at all steps of the advertising value chain connecting advertisers with publishers (e.g., ad servers, demand and supply side platforms, tracking and analytics) and is itself a publisher (e.g., through YouTube).

Being both an intermediary and a provider in the downstream market allows such vertically integrated platforms to attain data on other providers and businesses in the downstream market, including their customer. This presents a similar dilemma as in the case of ancillary services. Because the platform holds a gatekeeper position in one of the services (e.g., as a search engine, as an e-commerce marketplace platform, or as an ad server) third-party providers are forced to use the platform to make business or receive consumers' attention. But in doing so, **data is revealed to**

[249] Identity management services are not only offered by digital platforms, but also by other consortia, typically evolving around banks, insurance companies and telecommunications providers, like "Itsme". (https://www.itsme.eu) in Belgium, "BankID" (https://www.bankid.com) in Norway, "Verimi" (https://www.verime.de) in Germany or "Mobile Connect" (https://mobileconnect.io). However, to date, these play only a minor role for authentication to Internet services.
[250] Preibusch, S., Peetz, T., Acar, G., & Berendt, B. (2016). Shopping for privacy: Purchase details leaked to PayPal. *Electronic Commerce Research and Applications*, *15*, 52-64.
[251] Krämer, J., Schnurr, D., & Wohlfarth, M. (2019). Winners, losers, and facebook: The role of social logins in the online advertising ecosystem. *Management Science*, *65*(4), 1678-1699.
For a non-technical version see Krämer, J., Schnurr, D., & Wohlfarth, M. (2019). Trapped in the Data-Sharing Dilemma. *MIT Sloan Management Review*, *60*(2), 22-23.

the platform, who is at the same time a competitor in the downstream market. This decreases the ability of independent providers to compete against the data-rich platform. For example, the platform learns about popular product searches, consumer preferences and successful designs, which it can then use to improve its downstream services and products, incorporate innovative ideas and designs, or to maintain an advantage over collecting consumer profiles (e.g. for the use of advertising) over others. This is especially problematic as the platform has not only a data-driven advantage but can also use its role as an intermediary to steer consumers' attention to its products and services rather than to that of the independent providers which use the platform.

Allegations of this sort have been made, for example, against Amazon, who allegedly used data from independent sellers in Amazon's marketplace to launch competing products under its brand name (Mattioli 2020)[252] and then used its intermediation power to make its products more prominent than that of the competitors. Following such complaints, in July 2019 the European Commission has opened a formal antitrust investigation against Amazon regarding these practices.[253]

While such data advantages put third-parties at disadvantage and deny them scale to grow on the platform, one may argue that **consumers are not harmed in the short-run, because they eventually gain access to the innovative products** and services anyway, just through the platform itself rather than the third-party. In the short-run, the harm lies rather in the possibly unfair distribution of innovation rents between the third-parties and the platform. However, along the same lines as argued in Section 4.1.1, this threat of disintermediation by the platform will **likely stifle innovation and investment by third-parties in the long-run**, which does imply harm to consumers in the form of foregone innovations. We also note that this long-run harm is difficult to assess empirically because it would require to somehow re-construct the missing counterfactual (i.e., a world in which innovations were not copied by the platform).

This is corroborated by the empirical study by Zhu and Liu (2018)[254], who analyse data on 163,853 products sold on Amazon.com in 22 subcategories. On the one hand, the authors find that Amazon is more likely to enter product categories with an average customer rating of more than three (of five) stars and in which there already exist a relatively large number of independent sellers, among other things. However, following Amazon's entry, there is no significant increase (nor decrease) in customer satisfaction. Thus, one may conclude that from a customer perspective, Amazon's product does not improve over the original product. Rather, this suggests that entry leads to a shift in innovation rents from the third-party sellers to Amazon. On the other hand, and this is especially problematic from a dynamic welfare perspective, third-party sellers affected from entry are more likely than unaffected sellers to reduce the numbers of products they offer on Amazon, and some exit altogether. This effect is more pronounced for small sellers. This suggests that innovation and competition by third-party sellers are reduced following Amazon's entry.

Similarly, Wen and Zhu (2019)[255] study mobile app markets on Android and find that the mere threat of entry by Google in some mobile app markets led independent developers in those markets to reduce innovation and to raise prices of their affected apps. This suggests that after the platform owners entry, the independent developers pursue an exploitative strategy (i.e., extract surplus from their existing customer base as much as possible) rather than to compete with the platform owner. They also find, however, that subsequently the independent developers' attention is shifted to unaffected and new apps. This would mean that innovation incentives are not completely stifled following due to the threat of entry by the platform-owner. Nevertheless, the threat of entry reduces the prospective rents from innovation.

4.1.5 Kill zones and the impact on innovation and venture capital

Against the backdrop of the preceding discussion on data-driven network effects and venturing of data-rich firms into related and emerging markets, recent and more nascent literature has looked at the effects on financing for start-ups. According to these studies, there is growing empirical evidence

[252] Mattioli, D. (2020). Amazon scooped up data from its own sellers to launch competing products. The Wall Street Journal, April 23.
[253] https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4291
[254] Zhu, F., & Liu, Q. (2018). Competing with complementors: An empirical look at Amazon. com. Strategic Management Journal, 39(10), 2618-2642.
[255] Wen, W., & Zhu, F. (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. Strategic Management Journal, 40(9), 1336-1367.

that **some firms may have established 'kill-zones' around their core business model** (see, e.g., Scott Morton et al. 2019[256] for a thorough discussion, but also related news reports[257]). This means that innovative start-ups, which are either pioneering a technology or a market that may allow them to eventually develop data-driven network effects on their own, are typically facing one of two options. Either, the start-up is being bought by the data-rich incumbent; or the start-up must fear that the data-rich incumbent soon incorporates the innovation in its own service, making use of its lower cost of innovation due to data-driven network effects, and its larger installed base of users. In both cases, the start-up eventually vanishes as a potential competitor. This is what has been coined the 'kill zone'.

The first issue is being scrutinized more and more in merger review, and several policy proposals have been made to address it (see, e.g., Bourreau and de Streel, 2020[258]; Crémer, de Montjoye and Schweitzer 2019; Motta and Peitz 2020[259]; Scott-Morton et al. 2018), especially by lowering the threshold for merger review in data-driven markets, and by adopting a more dynamic viewpoint on the importance of nascent related markets for the contestability of an entrenched data-rich incumbent (see also Section 4.1.5.).

The second issue may be even more difficult to address. Data-rich incumbents differ from start-ups not only with respect to the already larger existing base of users, larger data-driven network effects and thus a lower cost of innovation and entry. They also have better access to financing and often deep pockets that make them largely independent of the need for venture capital. In reverse, the 'kill zones' seem to affect the venture capital market. In particular, start-ups that complement the incumbent's business model are more likely to receive venture capital than start-ups that challenge the incumbent (for a discussion see, e.g., Smith, 2018[260] and Rinehardt, 2018[261]). If this pattern is corroborated by further research, this would mean that market entry barriers are even higher and contestability becomes even more difficult than the preceding discussion has acknowledged.

4.1.6 Data-driven network effects and efficiency

While we have focussed on the data-driven theories of harm in the above discussion, we shall also explicitly mention the **inherent efficiency advantages that come along with data-driven network effects**. These arise precisely because of the same reasons, i.e., from data-driven economies of scale and scope and lower marginal costs of innovation. This makes any economic regulation in this context highly complex because one must carefully **weigh the efficiency gains stemming from competition and innovation from third-parties against the efficiency gains stemming from data-driven economies of scale and scope by the data-rich incumbent**. As discussed in the previous section, economies of scale (e.g., data about more users or 'broad data') and economies of scope (e.g., more heterogeneous data about users or 'deep data') allow firms to make better recommendations and to offer better fitting products, services and content, among other things. Also, Martens (2020)[262] notes that users can benefit from the aggregated view that a platform operator collects on supply and demand. Realizing economies of scale and scope in such data aggregation benefits consumers because it would be impossible for an individual consumer to collect this information herself and consequently not the same levels of efficiency could be achieved. Consequently, Martens (2020) concludes that data-driven network effects and "platforms are both a blessing and a curse in the digital data economy. [Platforms] are necessary intermediaries to

[256] Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Melamed, D. & Morgenstern, J. (2019). Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report. Draft. Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business.
[257] See, for example, The Economist (2018). Into the danger zone
American tech giants are making life tough for startups. Available at:
https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups; Financial Post (2018). Inside the kill zone: Big Tech makes life miserable for some startups, but others embrace its power. Available at: https://business.financialpost.com/technology/inside-the-kill-zone-big-tech-makes-life-miserable-for-some-startups-but-others-embrace-its-power
[258] M. Bourreau and A. de Streel, Big Tech Acquisitions: Competition & Innovation Effects and EU Merger Control, CERRE Issue Paper, February 2020 available at https://www.cerre.eu/publications/big-tech-acquisitions-competition-and-innovation-effects-eu-merger-control
[259] Motta, M. and M. Peitz (2020). "Big Tech Mergers," CEPR Discussion Paper 14353, available at https://cepr.org/content/free-dp-download-31-january-2020-competitive-effects-big-tech-mergers-and-implications
[260] Smith, N. (2018). Big Tech Sets Up a 'Kill Zone' for Industry Upstarts. Available at https://www.bloomberg.com/opinion/articles/2018-11-07/big-tech-sets-up-a-kill-zone-for-industry-upstarts;
[261] Rinehardt, W. (2018). Is there a kill zone in tech? Available at: https://techliberation.com/2018/11/07/is-there-a-kill-zone-in-tech/
[262] Martens, B. (2020). Data access, consumer interests and social welfare An economic perspective on data

generate benefits from data aggregation, realise data-driven positive network externalities and enable the emergence of new markets that were not feasible before the arrival of digital data. At the same time, data aggregation generates new sources of market failures that did not exist in the pre-digital economy."

4.2 The need for ex-ante regulation and its policy objectives

4.2.1 The requirement for ex-ante regulation

The EU policy debate on 'access to data' as a means to establish competitive and contestable digital markets has, rightfully so, departed from and centred around the possibilities under the existing competition law framework. Under this framework, before access to a dominant competitor's data set can be granted, a key criterion under the notion of the "essential facilities doctrine" is whether that data set is indispensable to compete in the same market and that it cannot be acquired in another way (Cremer et al. 2019, pp. 101). The possibility and limits of competition law in this context will also be discussed in more detail in the companion CERRE report by Feasey and de Streel (2020). Ultimately, indispensability is an empirical criterion that needs to be addressed on a case-by-case basis. However, corroborating the insights of many previous analyses (inter alia Graef et al, 2015[263]; Cremer et al. 2019) our case studies from Section 0 also show that such **'essential data' in the narrow sense often does not exist in many markets**, not even in the home markets of tech giants like Amazon, Google or Facebook. In search, only the web index is 'essential', which requires crawling publicly available websites. Setting up an e-commerce site is, of course, also possible without product recommendations. Likewise, a media platform just needs content and a delivery platform to launch but does not require user data to start, even if it pursues a (contextual) advertisement-based business model.

Therefore, it is a common understanding by now that the **role of competition law in granting access to user raw data is currently limited**. This is not merely due to the high legal standards that have to be met pursuant claims under the essential facilities doctrine, but also due to procedural issues. Competition cases take a long time, whereas potential competitors need to seek business opportunities quickly in the dynamic environment of digital markets (see Section 4.1). It is then difficult, if not impossible to develop remedies ex-post that would restore the market conditions like they were before the abuse or refusal to deal. Moreover, competition policy is not well suited to design complex remedies, especially if those would require ongoing monitoring for compliance (see, e.g., de la Mano & Padilla, 2018[264], Feasey & Krämer, 2019[265]). Nevertheless, ex-ante regulation also bears some inherent costs, which need to be taken into account.

Indeed, the **case studies have also provided a more detailed insight in what is necessary to be competitive in these markets, and this goes far beyond the minimum requirements of 'essential data'**. The case studies show that more data will gradually improve the quality of the service offered in these markets, albeit at a decreasing marginal rate. In search, ranking quality improves with the breadth and depth of the search logs available. Each search increases the observations per query over all users (breadth of the data), and, at the same time, the observations per user over all queries (depth of the data). This combination is what makes data valuable for similarity assessments that facilitate algorithmic learning, also for new users and new queries. In e-commerce, similar arguments hold for improving the quality of the recommender system and the quality of predicting demand. In media and advertising, likewise, the combination of breadth and depth of data allows media platforms to gradually improve content recommendation, content curation and the effectiveness of advertising, as well as to heighten their ability to demonstrate value and engage effectively in trading their inventory.
Thus, **while competition law continues to play an important role, it needs to be complemented with ex-ante regulation**, especially if potentially complex data access remedies are required that must balance competition with privacy rights. Moreover, as we will argue next, the policy objective should be to facilitate entry in related and niche markets, rather than to enable firms to take on data-rich incumbents head on in their home market. In a similar spirit, Cremer et al (2019,

[263] Graef, I., Wahyuningtyas, S. Y., & Valcke, P. (2015). Assessing data access issues in online platforms. *Telecommunications policy*, *39*(5), 375-387.

[264] De la Mano, M., & Padilla, J. (2018). Big Tech Banking. *Journal of Competition Law & Economics*, *14*(4), 494-526.
[265] Feasey, R., & Krämer, J. (2019). *Implementing effective remedies for anti-competitive intermediation bias on vertically integrated platforms*. Centre on Regulation in Europe (CERRE). Available at: https://www.cerre.eu/publications/implementing-effective-remedies-anti-competitive-intermediation-bias-vertically

p. 9) conclude that "Article 102 TFEU is not the best tool to deal with data requests by claimants who pursue business purposes that are essentially unrelated to the market served by the dominant firm (i.e. access to data for training AI algorithms for unrelated purposes); in such cases, the **emergence of market-based solutions or the adoption of a regulatory regime would seem preferable**" and "where a dominant firm is required to grant access to continuous data (i.e. to ensure data interoperability), there may be a need for regulation."

4.2.2 Contestability vs. niche entry and growth as the policy objective

In policy circles, including this report, it is often suggested that digital markets should remain 'contestable'. From a narrow economic perspective, the **theory of contestable markets** (Baumol, Panzar & Willig,1982)[266] contends that there may be markets in which, albeit only one firm can operate profitably, the monopolist will nevertheless behave as if it were in (perfect) competition. More precisely, the theory argues that, if the monopolist were to exercise its market power, say by raising the price above the competitive price, this would immediately lead to the entry of a new competitor, who could poach all of the monopolist's customers by offering a slightly lower price. In consequence, the fear of entry disciplines the incumbent monopolist restores competitive outcomes and makes regulation obsolete. The theory of contestable markets **rests on several assumptions that have proven to be difficult to materialise in practice**. In particular, two of the main assumptions are that (i) there are no sunk costs creating barriers to entry or exit, and (ii) that potential competitors are ready to step in the market at any time. These conditions are violated in traditional infrastructure markets, for example. Indeed, this failure to meet the criteria of the contestable market theory has been one of the main rationales for ex-ante regulation in infrastructure industries.

However, **contestability is a policy objective that is often associated with digital markets**. Possibly so, because it has been argued that "competition is just one click away", and consumers can easily multi-home between digital services, especially if they are provided at zero monetary price, which should facilitate switching. Moreover, it has been argued that hit-and-run entry may, after all, be possible, especially by other tech companies, who already have access to skilled labour and data centres, which they could readily repurpose. But also sunk costs for de-novo start-ups, so could be argued, are comparably low because the necessary technical infrastructure can be scaled dynamically or leased, e.g., by outsourcing computing infrastructure to the cloud. Following this view, at least some digital markets are contestable. Consequently, in such markets entry would be likely if the established firm tries to exploit its market power, e.g., by raising the price of a previously free service, degrading the quality of the service, displaying more advertisements or undermining privacy. The conclusion then would need to be that there is no scope for policy interventions. The highly dynamic environment of digital markets would ensure the necessity to maintain a continuously high innovative performance as a precondition to remain successful and to avoid "creative destruction" according to Schumpeter (see, e.g., Monopolkommission, 2015[267]).

However, as we have argued above, **in markets where data is a key input, the preconditions for contestability in the narrow economic sense are likely to fail**, because such markets are characterized by data-driven network effects, which can constitute significant entry barriers. Moreover, digital services often exhibit other significant direct or indirect network effects and are bundled with other services in a larger digital ecosystem, which violates the assumptions of the contestable market theory further. This is not to say that contestability is a myth in digital markets generally, but it is not a policy goal that seems realistic for data-driven markets, especially those surveyed in our case studies. It also does **not seem desirable to "replace" incumbents by another firm that fulfils essentially the same service** if this requires to duplicate the massive technical infrastructures necessary to provide the service, also from an environmental perspective.

Nevertheless, in a series of articles, Argenton and Prüfer (2012)[268], Prüfer and Schottmüller (2017)[269] and Prüfer (2020)[270] argue that contestability could theoretically be achieved in markets with data-

[266] Baumol, W., Panzar, J., & Willig, R. (1982). Contestable markets and the theory of industry structure. San Diego: Harcourt Brace Jovanovich.
[267] Monopolkommission. (2015). Competition policy: The challenge of digital markets. Special Report 68. Retreived from http://www.monopolkommission.de/images/PDF/ SG/s68_fulltext_eng.pdf.
[268] Argenton, C., & Prüfer, J. (2012). Search engine competition with network externalities. *Journal of Competition Law and Economics*, *8*(1), 73-105.
[269] Prufer, J., & Schottmüller, C. (2017). Competing with big data. *Tilburg Law School Research Paper*, (06).
[270] Prüfer, J. (2020). Competition Policy and Data Sharing on Data-driven Markets: Steps Towards Legal Implementation.

driven network effects by granting competitors access to the raw user data (especially observed, behavioural data) that fuel these effects. We remain **sceptical that contestability in a narrow sense can indeed be achieved**, for several reasons. Firstly, although sharing some data may be feasible and desirable, there are many caveats, which limit both the depth and the width of data which can be shared. Thus, entrants will never have access to the raw user data to the same extent as the incumbent. Secondly, even if sharing the full data set was possible, it also seems very difficult to verify for a competent authority whether indeed all data were shared. Thirdly, as argued before, entry barriers will almost certainly not only be constituted by data-driven network effects, but also by other means such as user-driven network effects (direct and indirect) and associated consumer lock-in as well as access to capital and complementary assets such as skilled labour.

Nonetheless, we also argue in this report that facilitating access to user data is an important building block in a policy framework for the digital economy, and in Section 0 we discuss the various options how this can be achieved alongside with their trade-offs. However, what is important to highlight here is that facilitating access to user data should not be pursued in order enable duplication of the services of existing dominant data-rich firms, but because it **stimulates niche entry and allows less data-rich firms to grow and scale**. This could include firms entering either existing or emerging markets, which are not yet dominated by a data-rich firm, carving out niches in existing markets through the development of new technology, or firms entering from adjacent markets. Firms can then slowly scale in this niche, develop data-driven network effects on their own, and venture into related markets to eventually become a sizeable competitor, which can potentially even exert competitive pressure on incumbent data-rich firms.

Niche entry is also the mode of entry that today's tech giants used. For example, Google gained a foothold in the market for search by offering a superior search technology, at a time where some "web portals" still maintained "catalogues" of the most important web pages and ranking was done mainly based on word frequencies. The web contained less than 300 thousand websites when Brin and Page launched the first version of their search engine named "BackRub" in 1996, and there were less than 3 million websites when Google was officially launched in 1998. Today there are more than 1.7 billion websites. Similarly, Amazon started out in 1995 as a pure online book retailer, one of the first of its kind. Books were chosen as the product category on purpose, because of the combination of low unit prices, high demand for literature and a large number of titles available made it ideal for an online business. The combination was also ideal to embrace relatively new technology at the time, personal recommendation systems. Amazon employed this technology early on and significantly developed it further (see Linden et al., 2003), which gave it a competitive edge over brick-and-mortar bookstores which did typically not offer such personalised recommendations. By comparison, Facebook was a latecomer and launched in 2004. However, it also chose a niche entry, by exclusively addressing college students first, offering isolated versions of "Facebook" for each school before merging into a single platform. This entry strategy helped it to overcome the network effects of the then-dominant social network MySpace. However, social networking was still in its infancy. MySpace had been launched one year earlier in 2003 and had quickly superseded the first social network Friendster. At its peak in 2008, MySpace had fewer than 80 million unique visitors/month. By comparison, Facebook currently has about 2.6 billion visitors/months, and 1.8 billion visitors per day. Moreover, as Parker et al (2016)[271] lay out, MySpace subsequently also suffered from issues related to content quality and management, which eventually allowed Facebook to supersede it in size and relevance. Facebook was ad-free until November 2007.

This underscores that enabling more niche entry and niche growth should be the primary focus of public policy. To this end, in Section 5, we explore various possible remedies that facilitate or limit data access and discuss how this may support niche entry and growth. Again, this does not mean that existing data-rich incumbents will or should vanish. They have mastered specific markets and achieve significant efficiency in doing so. At the same time, this does not mean that such dominant firms should receive a regulatory free pass. In some cases, especially in the case of vertically integrated platforms that have a "significant market status" (Furman et al., 2019[272]) or are an "unavoidable trading partner" (Cremer et al, 2019) **additional regulation to ensure fair and transparent competition on the platform may be warranted**. This should be done primarily

[271] Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy? and How to Make Them Work for You*. WW Norton & Company.
[272] Furman, J., Coyle, D., Fletcher, A., McAules, D., & Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. *Report prepared for the Government of the United Kingdom, March*.

with a focus on enabling business users to grow on the platform and to achieve significance in a narrow market segment first. A comprehensive policy to build viable competitors in new markets will likely also include additional safeguards to merger policy and the 'kill zones' mentioned in Section 4.1.5. As far as policy measures and regulation go beyond the access to data, they are beyond the scope of this report. Facilitating more access to data will, however, likely be a key element of future policy interventions in the context of digital markets.[273]

Unless there is significant mismanagement, we argue that **only disruptive innovation (e.g., the rise of mobile communication and smartphones or artificial intelligence) or significant changes to consumer preferences (e.g., for privacy concerns) or other external shocks (e.g., a pandemic forcing people to use new services) may indeed lead to a Schumpeterian 'creative destruction' of a tipped market**. Besides, this is only likely if such disruption is combined with effective policy measures that protect innovative newcomers from being 'killed'. Niche entry, however, can occur at all times and does not require disruption per se. Ultimately, as new markets and services emerge and some may vanish, the future digital markets landscape as a whole should be characterised by a large diversity of players, all of whom receive a significant share of consumers' attention and data. As data-driven economies of scale and scope tend to have decreasing marginal returns for large data sets (see Section 3.2), all of these players would have the capability to compete and innovate at high levels of efficiency. Existing incumbents will probably continue to master specific markets due to the economic forces laid out above. However**, if the digital competition landscape is more diverse, competition *for* new markets would then be more intense and occur on a more level playing field**; and as long as policy measures such as data sharing and merger control are effective, entry and growth by newcomers would be possible. This is the nature of contestability in the broad sense that we envision as the policy goal for this report.

4.2.3 How can the effectiveness of a data access policy be evaluated?

Consequently, a **crude assessment of whether the policy objective is achieved could be to measure the concentration and distribution of independent providers based on usage statistics**, such as the *daily time spent using a site or service* or *daily views/sessions per visitor*, or preferably a combination of these. This would be a measure of the level of competition for user's attention and give an idea of the extent to which various players, and potentially any niche competitors, have captured shares of the attention market.

SITE RANK	SITE	DAILY TIME ON SITE	DAILY VIEWS PER VISITOR	% OF TRAFFIC FROM SEARCH	TOTAL SITES LINKING IN
1	Google.com	13:46	15.21	0.40%	1,947,950
2	Youtube.com	13:35	7.63	15.90%	1,494,536
3	Google.be	2:42	3.60	4.90%	7,542
4	Facebook.com	18:21	8.07	8.10%	3,373,532
5	Reddit.com	5:52	4.51	27.20%	238,198
6	Wikipedia.org	3:56	2.95	71.40%	1,194,143
7	Live.com	5:00	5.16	11.90%	38,849

Figure 4: Alexa web usage statistics for Belgium, April 2020, ranked according to the Alexa traffic rank (a combination of daily visitors and page views)[274]

To exemplify this point, consider the Internet usage statistics that are provided by a company named Alexa[275] (which was acquired by Amazon already in 1999 for $250 million). Figure 4 shows a snapshot of the usage statistics that Alexa gathers, in this case for Belgium for April 2020.[276] The statistics include measures of visitors' daily time on site and the daily pageviews per visitor, which are both used to measure how much these websites can keep consumers engaged. But they also include measures of dependency on others, such as how much of the traffic originates from search. Finally, the statistics also include measures of relevance, such as the number of total sites linking in on that website. This roughly resembles the PageRank relevance measure used by Google as one of the signals to rank search results (see Section 2.1). The usage data gathered by Alexa is much richer

[273] See, for example, Prager, A. (2019). Vestager calls for more access to data for smaller platforms. Euractiv. Available at: https://www.euractiv.com/section/data-protection/news/vestager-calls-for-more-access-to-data-for-small-platforms/
[274] https://www.alexa.com/topsites/countries/BE
[275] http://www.alexa.com
[276] It should be noted, however, that these statistics include only usage of the Internet through websites, and not through other services (e.g., apps).

than this snapshot, but it is already easy to see how a measure of the concentration of users' attention can be devised, and how, at the same time, it can be controlled for by measures of dependency and relevance. As the distribution of users' attention is going to be highly skewed, following a long-tail distribution, an appropriate concentration measure may be constructed more similar to a Gini coefficient rather than to a Herfindahl index.

While we intend to stimulate a debate on the appropriate measurement rod of economic policy interventions in digital markets, any such measurement would need significantly more deliberation and investigation than what we have laid out here, of course. Several other indicators specific to types of services might be used as well. For example, **media regulators have several ways** that **measure diversity in media markets**. Consumer expenditure patterns might indicate a diversity of use for e-commerce. Policy discussions on contestability of platform markets need to give due attention to devising ways to assess whether or not necessary conditions for niche entry and growth have been achieved for various contexts.

05

POSSIBLE DATA ACCESS REMEDIES AND THEIR ECONOMIC TRADE-OFFS

5 Possible data access remedies and their economic trade-offs

We now turn to the discussion of potential market interventions, specifically data access obligations, in light of our previous insights on the value of data and the need for ex-ante regulation. We will focus on a more economic discussion, and defer for their legal implementation to the companion CERRE report by de Streel and Feasey (2020)[277].

5.1 Possible data remedies that limit the collection of user data

In this section, we discuss possible policy measures that aim at **levelling the playing field between data-rich incumbents and (potential) competitors by limiting the ability of data incumbents to collect ever more user data and to combine the data sets derived from different services**. Levelling the playing field in new and emerging markets would be an essential step to enabling niche entry and effective contestation in the long run. We consider several proposals for data remedies that have been made in this context and point to the economic trade-offs involved if those remedies were to be implemented.

5.1.1 Data Silos / Chinese data walls

As detailed above, there can be significant value in combining data sets from different services to derive deeper and more precise user profiles. A **defining feature of many data-rich firms in the digital economy is that they can collect user data from various sources** (see Section 3.1). User data is usually not only collected directly from the consumer-facing web services (e.g., search, mail, maps, video and audio streaming, online shop) but also from complementary software (e.g., browsers, operating systems, apps) and devices (e.g., voice assistants, video streaming devices, smartphones and tablets). Data-rich firms typically have multiple consumer-facing services and products through which they can collect user data, not the least due to the "domino effect" highlighted in Section 4.1.2 and as a result of previous mergers and acquisitions. However, as discussed in Sections 3.1.2.4 and 4.1.3 user data is also collected from a potentially large number of third-party service providers through ancillary data services (e.g., web analytics, payment or identity solutions) that data-rich firms offer.

In an effort to level the playing field for new and emerging markets, a potential remedy could be to **limit domino effects by constraining the data-rich firm's ability to combine user data originating from various services and data sources**. Instead, the data should be kept in separate databases or data silos pertaining to the service where they were initially collected. In theory, this could provide a more level playing field and therefore encourage entry into niche markets, because both data-rich and data-poor firms would first have to develop data-driven economies of scope in the new market.

However, almost by definition, **such 'Chinese data walls' also limit the inherent efficiency advantages** that come along with data-driven economies of scale and scope (see Section 4.1.6), both in the future and, if applied to existing services, also for the current services ecosystem. For example, data interoperability between an e-mail service, a calendar service and a map service has evident advantages for consumers, because appointments can be readily created from an incoming e-mail and travelling time and directions can be readily derived from the map service. But in other cases, such synergies between the data collected in various services may not be apparent to consumers or even be detrimental. This is particularly likely if consumers' privacy is undermined, e.g., by tracking them across the web, for the main purpose of targeted advertising.

In any case, this type of remedy seems **difficult to monitor and to enforce** for at least two reasons. Firstly, there is an inherent **information asymmetry between the regulator and the regulated firm**. It will generally be very difficult to detect and to prove where the data about a consumer originated and whether or not data sets were combined. At the same time, combining data sets from various services and harnessing data-driven network effects will often be too tempting for firms in light of this information asymmetry. As a prominent example, Facebook announced in relation to its 2014 acquisition of WhatsApp multiple times to the European Commission that it did

[277] De Streel, A., & Feasey, R. (2020). Data Sharing for Digital Markets Contestability: Towards a Governance Framework. CERRE Policy Report

not intend to match user profiles from WhatsApp with those of Facebook, and that it would not be able to do so. However, it later became clear that Facebook was already in 2014 aware of the possibility that such user profiles could be matched, but this became known to authorities only after Facebook publicly announced the linking of the profiles in August 2016 in its terms of service and privacy policies. This led to a fine of €110 million for providing incorrect or misleading information to the European Commission in 2017.[278] So, while some level of public monitoring is possible through the privacy policies, these are, of course, provided by the firms themselves.

Secondly, there seems to be a **trade-off between competition-related policy measures and privacy regulation**. This is prominently exemplified in the Bundeskartellamt vs. Facebook case, where, for the first time, an authority has ordered data siloing by default as a remedy in a competition case. Specifically, Facebook was ordered to separate the data that it collects about users on Facebook itself, and the data that is collected on other sites (e.g. Instagram and WhatsApp, but also third-party websites and apps). According to the decision, consumers would need to explicitly consent to the combination of data from the various sources (opt-in) in the future.[279] Several legal scholars question whether the Bundeskartellamt, as a competition authority, has overstepped its legal mandate because the ruling relates to a privacy issue rather than a competition issue. This highlights one possible legal tension that can arise when policing user data as a competitive factor. It remains to be seen how this case proceeds further. Facebook's appeal of the ruling was at first successful at the Düsseldorf Higher Regional Court, which found in a preliminary ruling that the combination of data sources did not constitute anti-competitive effects.[280] This suspended the enforcement of the Bundeskartellamt's order. However, on June 23, 2020, the German Federal Court of Justice (Bundesgerichtshof), also in a preliminary ruling, found that the Bundeskartellamt's order against Facebook can be enforced, overruling the Düsseldorf Higher Regional Court's decision.[281] In the next step, the Düsseldorf Higher Regional Court has to take a final ruling on the case, after which likely the Federal Court of Justice is to decide again.

The Bundeskartellamt vs. Facebook case relates to a wider issue with respect to the trade-off between privacy and competition. **Several commentators criticize that the GDPR has, ironically, facilitated the dominant market position of data-rich** firms vis-à-vis smaller firms. Geradin, Karanikioti and Katsfis (2020)[282] present several arguments for this. Among other things, they mention GDPR compliance costs, the greater ability of dominant firms to acquire consumer consent and the hindrance of GDPR to facilitate data sharing. Their most relevant argument for the present context addresses the problematic role of the GDPR's one-stop-shop mechanism, whereby "the supervisory authority of the main establishment or the single establishment of the controller or processor shall be competent to act as a lead supervisory authority."[283] This, so the authors argue, has granted "disproportionate enforcement power" to certain national data protection authorities (DPAs), particularly to the Irish DPA, who is therefore responsible for some of the largest tech firms. At the same time, the Irish DPA has a reputation for being lean on enforcement and is, especially relative to its importance, understaffed and underfinanced.[284]

Against this backdrop, allegations are being made that data-rich incumbents do not comply with GDPR's principles of data minimization and purpose limitation. For example, Geradin et al. (2020) note that in 2012 Google consolidated more than 60 separate privacy policies into a single privacy policy, which allows Google to combine data it collects across its various consumer-facing services for a wide variety of purposes. At the same time, similar as in the case of Facebook, users only have

---

[278] https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1369
[279] Bundeskartellamt (2019). Bundeskartellamt prohibits Facebook from combining user data from different sources, Available at https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html
[280] Lomas N. (2019). Facebook succeeds in blocking German FCO's privacy-minded order against combining user data", Tech Crunch, Available at https://techcrunch.com/2019/08/26/facebook-succeeds-in-blocking-german-fcos-privacy-minded-order-against-combining-user-data/
[281] Bundesgerichtshof (2020). Bundesgerichtshof bestätigt vorläufig den Vorwurf der missbräuchlichen Ausnutzung einer marktbeherrschenden Stellung durch Facebook. Press Release. Available at: https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2020/2020080.html;jsessionid=F02FBF1A27F70DFD5DD8DC318EFB6C59.1_cid368?nn=10690868
[282] Geradin, D., Katsifis, D. and Karanikioti, T. (2020). GDPR Myopia: How a Well-Intended Regulation ended up Favoring Google in Ad Tech. TILEC Discussion Paper No. 2020-012. Available at: https://ssrn.com/abstract=3598130
[283] GDPR Article 56(1).
[284] Vinocur, N. (2019). 'We have a huge problem': European tech regulator despairs over lack of enforcement, Politico, 27 December 2019, Available at https://www.politico.com/news/2019/12/27/europe-gdpr-technology-regulation-089605; Satariano, A. Europe's Privacy Law Hasn't Shown Its Teeth, Frustrating Advocates. The New York Times, 27 April 2020, Availabe at: https://www.nytimes.com/2020/04/27/technology/GDPR-privacy-law-europe.html.

limited ability to opt-out of such a combination of data from various services. In a formal complaint filed to the Irish DPA, the niche web browser company Brave alleges that Google would use "hundreds of purposes to justify data processing activities", including – and in line with the domino-effect highlighted above – the explicit purpose to collect data to "help develop new [services]". [285] The result would be an "an internal data free-for-all" practice, whereby users' consent for one service is used to distribute that data freely "inside the black-box" and with an "unknowable number of external business partners" [286] while lacking transparency how and with whom data processing is taking place. As Condorelli and Padilla (2020)[287] highlight, this practice of "privacy policy tying" is also a source of data-driven envelopment (see Section 4.1.2). As described in Section 2.1 in the case study on web search, user data from other services are indeed a valuable input resource to improve the ranking quality of search results.

It is yet too early to know what the outcome of these complaints is going to be. In a related case regarding the obligation to create a user account when setting up an Android phone, the French DPA, CNIL, has already found that Google was lacking transparency about how data was used and was lacking "specific" and "unambiguous" consent for "ads personalization".[288] It fined Google €50 million for these violations.

In the future, more cases on the "internal data free-for-all" issue are likely going to be filed and decided. In contrast to Condorelli and Padilla (2020), we do **not believe that additional regulation to maintain data silos would be needed or that such obligations would be fruitful in light of the possible efficiency losses. Instead, the regulations set out by GDPR seem fit for purpose** and therefore DPAs are, in principle, legally equipped to address this issue. However, to provide a level playing field for data-rich incumbents and niche entrants, data conglomerates need to obtain valid and purpose-specific consent for each service that they offer, requiring users to opt-in, rather than to opt-out, of the combination of their data. This requires, first and foremost, **effective implementation and enforcement of the GDPR, and well-equipped DPAs** that have the resources to address and fine misuse promptly. In this context, it may be worthwhile to re-think the one-stop-shop mechanism, for example, by allowing the lead supervisory authority to agree with another national DPA on a delegation of the case to the other DPA. With such a mechanism the burden of the large and resourceful cases could be shared better among European DPAs, while not compromising on the appeal of a "one-stop-shop" mechanism. Alternatively, an EU authority (e.g., the Commission) could be put in charge in cases involving a large platform with pan-European systemic importance.

5.1.2 Shorter data retention periods

Another proposal to reduce the amount of personal data held by data-rich firms is to **limit the retention period for raw user input data** (e.g., search queries, clickstreams, location data, or other tracking data). Formally, firms are already obliged under GDPR to retain personal data only so long as it is necessary for the purposes for which it is used. Once this purpose is achieved, the data has to be deleted (Art. 6 GDPR). However, in practice, it is difficult to delineate the maximal necessary length of data retention. Pursuant to a right of access request (Art. 15 GDPR) or data portability request (Art. 20 GDPR) data subjects will often learn that personal data about them is stored for years, often dating back to the date when they first began using the service.[289] However, the purpose of this section is not to discuss the lawfulness of various firms' retention policies with respect to Art. 6 GDPR. Instead, we wish to discuss, from an economic point of view, the likely impact on market contestability and niche entry if the maximum retention period for personal data were much shorter, say three to six months, unless consumers explicitly opt-in to a longer period.

---

[285] Murgia, M. (2020). Google accused by rival of fundamental GDPR breaches. Financial Times. Available at: https://www.ft.com/content/66dbc3ba-848a-4206-8b97-27c0e384ff27

[286] See also Murgia, M. (2019). Google accused of secretly feeding personal data to advertisers. Financial Times. Available at: https://www.ft.com/content/e3e1697e-ce57-11e9-99a4-b5ded7a7fe3f

[287] Condorelli, D., & Padilla, J. (2020). Harnessing Platform Envelopment in the Digital World. Journal of Competition Law & Economics.

[288] https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc

[289] For example, personal data stored by Google can be accessed here: https://myactivity.google.com ; and personal data stored by Facebook can be accessed here: https://www.facebook.com/your_information/; Amazon does not offer a comparable dashboard where users can view all information stored about them, but they can, of course access their entire purchase history.

First, a shorter retention period would probably largely maintain a firm's ability to learn from data. For example, it is possible to apply incremental leaning techniques, which train a new model based on an existing one, and only use the new (incremental) data in the process. Nevertheless, some flexibility and efficiency will be lost, because it is not possible to go back and re-train a completely new model if the original data has been deleted.

Second, in reverse, this means that shorter data retention periods would rather benefit those firms that already have a large user base and therefore receive more data in a given period. By contrast, nascent firms with a relatively small user base may need to collect data over long periods of time in order to achieve a significant scale of data (with regard to both breadth and depth dimensions) at which data analytics becomes meaningful.

Third, shorter data retention periods would also limit the user data that could be shared with others, be it for commercial or regulatory reasons. For example, this would limit the amount of personal data that an individual could port to a new provider using its right to data portability (Art. 20 GDPR). This would then limit the new provider's ability to derive the same algorithmic insights from the data as the old provider, and more generally reduce the ability to re-use the data in other contexts. The trained algorithm would not be subject to data portability, as it represents 'inferred data' and not data that is 'provided' by the data subject.

Fourth, as a side note, we also mention that shorter data retention periods could also affect the appetite for data-driven mergers. Mergers that are mainly driven by the desire to combine the user data of the acquired firm with that of the acquiring firm are less attractive if the available user data is limited to a period of three to six months.

Taken together, an explicit limit to the default retention period for consumer data may be viewed positively from a privacy perspective.[290] Besides, it may not sacrifice too much efficiency with respect to algorithmic learning and reduce the risk of data agglomeration through mergers. However, it is **questionable whether shorter data retention periods would indeed achieve the main policy objective of increasing the competitiveness of third-parties and facilitating niche entry**. Our above reasoning shows that they would rather **benefit those firms that already have reached significant scale and increase market entry barriers** further.

5.1.3 Prohibit buying into defaults

It is well known in the behavioural economics literature that default settings have a powerful impact on consumer choice and nudge them towards a preference or bias for the default option. Generally, such nudges can be in the interest of consumers, but they can also be exploited to entrench dominant market structures. Consequently, **default settings, e.g., settings for the default search engine in browsers or pre-installed apps on Internet access devices, are very powerful instruments to secure consumers' demand which continues to fuel data-driven network effects**. Thus, firms are willing to pay large amounts to secure them the advantage of being the default option. For example, Google purportedly paid Apple $1 billion in 2014 to be the default search engine on Apple's browser Safari. Estimates are that this number has gone up significantly since then and is now between $9 billion to $12 billion.[291] Likewise, in its report on "online platforms and digital advertising" the Consumer Markets Authority (CMA) found that in the UK alone Google was willing to pay around £1 billion, which corresponds to 16% of all its search revenues, to be the default search engine on mobile devices.[292]

Hence, a straightforward policy proposal would be to **prohibit dominant firms (e.g., those with 'significant market status') to buy into such defaults**, particularly in the form of default settings in Internet access devices and the form of pre-installed software. Indeed, there have been prominent *ex-post* competition cases in this context, often coupled with issues of leveraging market power, such as in the cases involving Microsoft, which pre-installed its browser and media player on Windows, and Google, which made contractual arrangements to have its suite of apps pre-installed on Android. However, the policy proposal at hand would be to complement general ex-post oversight by constituting an *ex-ante* ban of buying into defaults in a *pre-defined number of settings* and *only*

[290] However, it may create some legal uncertainty for market players, because such a provision is not (yet) harmonised with the horizontal provisions under GDPR.
[291] https://searchengineland.com/report-google-to-pay-apple-9-billion-to-remain-default-search-engine-on-safari-306082
[292] https://www.gov.uk/government/news/cma-lifts-the-lid-on-digital-giants

for regulated firms (e.g., those with 'significant market status'). This is a remedy that, among others, the CMA seems to be sympathetic to following its investigation in digital advertising markets.[293]

An alternative, less interventionist, the approach would be to **mandate choice screens** (a proposed remedy in the cases involving Microsoft Windows and Google Android), or to auction off the default setting in a competitive procedure. However, we are sceptical as to whether this would indeed have a significant impact on the status quo. We discuss both options in turn.

Firstly, while **choice screens** are indeed a means to overcome the default bias and require consumers to make an active decision, they do not overcome the (data-driven) network effects that a dominant firm has already built-up. As explained above, data-driven network effects are to some degree a self-fulfilling prophecy. The firm that has accumulated the largest amount of consumer data does indeed offer the best service (e.g., search engine) because it has access to such data and others do not. Thus, consumers would largely self-select into the dominant service for this reason. In other words, even if an alternative provider would have (instantaneously) a better service if provided with the same amount of user data, consumers would not be able to realise this and would fail to coordinate to select this alternative provider individually. Moreover, in practice, the data-driven network effect does not materialize instantaneously, of course, which further exacerbates the coordination failure.

In reverse, this also means that **prohibiting dominant firms from being the default option** also bears inherent costs to consumers, because this increases the likelihood that consumers choose an inherently disadvantaged provider, which – by some objective standards – offers an inferior service, e.g., with respect to prediction accuracy. Of course, consumers are likely to have multi-dimensional preferences, e.g., regarding privacy, and the prediction accuracy of the search algorithm is arguably just one relevant dimension. But, as explained in Section 0, there is some trade-off between privacy (data minimisation) and algorithmic performance. By revealed preference, many consumers seem to opt for better algorithmic performance over privacy, which shows, e.g., in minimal market shares for privacy-preserving search engines such as DuckDuckGo. This dilemma between a possible loss of consumer welfare in the short-run (by diverting them from the 'best' algorithm) and possibly higher consumer welfare in the long-run (by facilitating competition and entry) cannot be overcome in the presence of data-driven network effects. Furthermore, as highlighted in Section 4.1, the replacement of the dominant provider in its home market and diminishing its efficiency may not be a realistic nor a desirable policy objective.

Secondly, **auction mechanisms are also not likely to achieve any other outcome than the status quo**. A dominant firm has the largest financial means and the largest incentive to protect its dominant position and would, therefore, win the auction anyhow.[294] Already today, where such auctions for defaults are not formally conducted, there is some competition for becoming the default. For example, in 2012 Mozilla's Firefox browser announced that it had reached an agreement with Yahoo as the default search engine option, replacing Google, which had been the default search engine on Firefox for a decade.[295] In 2017, Firefox switched back to Google as the default search engine in most countries outside China and Russia. Thus, there already exists a 'default tax' (in the sense of a necessary sharing of revenues) that dominant search engine providers pay to other providers with strong consumer-facing products, such as Apple and Mozilla. An auction would not change the market outcome significantly. Rather, by committing to an auction, the default-setting firm loses some strategic freedom, particularly the threat of a 'trigger strategy' to choose a default at its choosing. Thus, if anything, auctions are likely to reduce the 'default tax' that would be paid.

Finally, on a related note, prohibiting dominant firms from buying into defaults would also mean that other providers with strong consumer-facing products, such as browsers or other Internet access devices, would not be able to collect a 'default tax' of similar size. In other words, in addition to the possible loss in (short-term) efficiencies, there would also be less distribution of the benefits from efficiency (or wealth) from the dominant firm to other firms in the Internet economy. This seems especially problematic when independent non-profit firms like the Mozilla Foundation are on the receiving end of such wealth redistribution. Indeed, the lion's share of Mozilla's funding comes from

[293] See no 80, p. 25 in CMA (2019). Online platforms and digital advertising. Market study interim report. Available at: https://assets.publishing.service.gov.uk/media/5dfa0580ed915d0933009761/Interim_report.pdf
[294] See also CMA (2019), p. 83 for the same argument.
[295] https://www.zdnet.com/article/googles-back-its-firefoxs-default-search-engine-again-after-mozilla-ends-yahoo-deal/

selling defaults in the Firefox browser.[296] Limiting its ability to 'tax' dominant firms would also diminish its ability to provide an alternative and competitive consumer-facing product. Ultimately, this would drive consumers into vertically integrated alternatives, such as Safari or Chrome, thus increasing the concentration of firms with access to consumers' attention.

5.1.4 Line of Business Restrictions

One of the most interventionist policy proposals to limit the ability of data incumbents to collect and combine ever more user data is to limit the markets and services that firms with 'strategic market status' may venture and operate in. Such "Line of Business Restrictions" (LOBRs) may be applied prospectively and retrospectively and can involve either horizontal or vertical business restrictions (OECD, 2020).[297] Thus, in its most extensive form, such LOBRs amount to vertical or horizontal structural separation.

5.1.4.1 Pros and Cons of LBORs and structural separation in digital markets

LOBRs and separation regimes have been used in several (network) industries in the past, including energy, railroads, banking, television and telecommunications– often applied as a remedy when a margin squeeze has been identified as an abuse of dominance (OECD, 2020). However, in retrospect, **not all of these interventions have proven to be effective** and more recently the idea of structural separation has not been very popular among policy makers – neither in an ex-post nor in an ex-ante context. Nevertheless, the idea of LOBRs and structural separation in the context of digital markets has been floating in policy circles for some years and has been popularized more recently in 2019 in the US by legal scholars Lina Kahn[298] and Tim Wu,[299] and Senator Warren's[300] proposal to "break up big tech", especially in vertical relationships where a firm operates both the marketplace and acts as a downstream competitor on that same marketplace. Indeed, building on its policy to protect small domestic businesses from foreign direct investment, India already introduced a new law in February 2019 which bans foreign-owned e-commerce platforms to sell directly to consumers, forcing Amazon and Flipkart (a prominent Indian marketplace whose majority owner is Walmart) to change their local business practices.[301]

Proponents of LOBRs argue that only structural separation can truly **resolve some of the incentive issues that arise when operating at different levels of the value chain**. This relates, for example, to the incentive issues raised against Amazon[302] and Apple[303] in the respective EU antitrust investigations, but structural separation has also been proposed as a remedy in the context of the Google Shopping case.[304] It is also being considered by the CMA as a remedy to address concerns relating to Google's conflicts of interest in the open display market, e.g., by separating the ad server (selection and pricing of ads) from Google's other commercial activities (CMA 2019, p. 26). LOBRs also achieve data siloing as an unavoidable side effect, while resolving the associated monitoring and enforcement problems (see Section 5.1.1). Moreover, Khan (2019) argues that LBORs may also address several other non-economic policy goals, such as promoting diversity, preserving system resiliency and administrability.[305]

However, most economists and policy makers are concerned with the **problems associated with LBORs and separation**. Firstly, there are several practical problems of unbundling, which are often summarized under the metaphor "unscrambling the scrambled eggs". In many cases it is difficult if not impossible to separate different digital markets or services and to delineate the boundaries between them. However, this is not always the case, of course, and should not be seen as a killer

---

[296] https://www.cnet.com/news/in-major-shift-firefox-to-use-yahoo-search-by-default-in-us/
[297] OECD(2020). Line of Business Restrictions - Background note. OECD Working Party No 2 on Competition and Regulation. DAF/COMP/WP2(2020)1
[298] Khan, L. M. (2019). The separation of platforms and commerce. Columbia Law Review, 119(4), 973-1098. Available at: https://columbialawreview.org/content/the-separation-of-platforms-and-commerce/
[299] Tim Wu, *The Curse of Bigness: Antitrust in the New Guidled Age*, Columbia Global Reports (New York: Columbia Global Reports, 2018).
[300] Warren, E. (2019). https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c
[301] Findlay, S. and Kazmin, A. (2019). India's ecommerce law forces Amazon and Flipkart to pull products. Financial Times. Available at: https://www.ft.com/content/29a96ff6-2615-11e9-8ce6-5db4543da632
[302] https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4291
[303] Toplensky, R. (2019). Brussels poised to probe Apple over Spotify's fees complaint. Available at: https://www.ft.com/content/1cc16026-6da7-11e9-80c7-60ee53e6681d
[304] Bershidsky, L. (2019). Breaking up big tech is too scary for Europe. Bloomberg. Available at: https://www.bloomberg.com/opinion/articles/2019-03-13/breaking-up-amazon-facebook-and-google-is-too-scary-for-europe
[305] This is mentioned for completeness. Non-economic rationales for regulation are outside the scope of this report.

argument. Secondly, and probably even more problematic from an efficiency perspective, the information generated in one market or service can have strong positive spillover effects in the other market. Indeed, this will often have been the driving motivation for venturing in that market in the first place (see Section 4.1.2). The associated inefficiencies have already been addressed in the context of data siloing in Section 5.1.1., while the inherent efficiencies of data-driven network effects have been discussed in Section 4.1.6. Thirdly, standard economic reasoning argues that separation generally diminishes economies of scale and scope, and vertical separation, in particular, bears inherent inefficiencies due to double marginalization. However, this has also not prevented policy makers in the past to apply LOBRs if the incentive problems and inherent inefficiencies that stem from a lack of competition and innovation were deemed larger than the aforementioned potential losses inefficiency. Therefore, in the following two subsections, we highlight some additional arguments for and against vertical separation that arise particularly in digital platform markets, where the platform acts both as an intermediary as well as a competitor on the platform.

5.1.4.2 Vertical separation when the platform is both intermediary and downstream competitor

Firstly, in the context of a vertically integrated platform De Cornière and Taylor (2019)[306] study in a game-theoretical model the impact of **'biased intermediation'** (i.e., steering consumers attention and demand to the own, vertically integrated product or service) on product and service investment and innovation, and on consumer welfare. Biased intermediation is one of the key incentive problems that may arise in the context of vertically integrated platforms and that can potentially be addressed by vertical separation. The authors' key policy message is that whether or not biased intermediation reduces consumers' welfare and is thus indeed a problem depends crucially on whether a seller's main strategic variable is the *price* (and quality differences between products or services offered on the platform are less important) or the *quality* (and price differences are less important).

In the first case, **where the price is the main strategic variable**, sellers' and consumers' incentives are misaligned, a scenario the authors call "conflicting payoffs". This is because the price just leads to a welfare transfer between a consumer and a seller but does not increase overall welfare. A higher price is good for the seller, but bad for the consumer, and vice versa. In the second case, where quality is the main strategic variable, sellers' and consumers' incentives are aligned. This is because a higher quality increases welfare, and both the seller and the consumers can appropriate part of the welfare gain. This is what the authors call "congruent payoffs".

In the environment where payoffs are conflicting (price-driven platform markets), biased intermediation is indeed harmful to consumers, because it tends to steer consumers to 'bad offers' with high prices. In this case, and only in this case, vertical separation can indeed increase consumer welfare. However, in the case where payoffs are congruent (quality-driven markets), bias intermediation can have positive effects and vertical separation tends to *reduce* consumer welfare. This is because biased intermediation allows the vertically integrated firm to attain a larger demand for its product or service, which provides it with an additional incentive to improve the quality. The assumption here is that investment in quality is largely due to fixed costs (which is always the case for digital goods, such as software and apps). More demand then translates into larger economies of scale. Thus, similar to data-driven network effects, an important positive feedback loop is at work here. The seller that is promoted on a platform (even though it may initially not offer the 'best' product or service for consumers), has subsequently the highest incentive to improve the quality of its product or service (due to increased economies of scale), and is likely to eventually be the best (unbiased) choice for consumers. In a dynamic context, this has important ramifications for the burden of proof in ex-post competition cases that try to confirm, sometimes years later, whether or not intermediation was 'biased'. If this assessment is done too late, and intentionally 'biased intermediation' may have transformed into 'unbiased intermediation'. In any case, vertical separation would introduce competition for prominence among sellers and allow the platform to auction off the most prominent position to the highest bidder. In this way, the platform extracts some revenue from the prominent seller, which reduces its incentive to invest in quality. Consequently, when firms compete in quality, vertical separation leads to lower quality investments and thus lower consumer welfare. The authors also show that an environment in which firms compete in quality *and* prices, the results are similar to the case of "congruent payoffs", i.e., where firms compete rather in qualities. This underscores the importance of this setting and the potential loss of welfare from vertical separation.

[306] De Cornière, A., & Taylor, G. (2019). A model of biased intermediation. The RAND Journal of Economics, 50(4), 854-882.

Secondly, Krämer and Zierke (2020)[307] study platforms that host free-to-view content and demand a revenue share from the independent content providers. Examples are media platforms, such as Youtube, where the platform takes a share of the advertising revenue generated by independent content, and free apps in mobile app stores, where the platform takes either a share of the advertising revenues or a share of the in-app purchases. Content providers compete for customers on the platform based on the quality of the content (i.e., a "congruent payoffs" environment is considered where quality is the main strategic variable) and through prominence on the platform (e.g., in a search results list). Content providers can be of high or low efficiency with their ability to produce content quality. The authors study, among other things, the case of vertical integration where the platform always favours its content on the platform. Similar to what was found by De Cornière and Taylor (2019), they highlight that **vertical integration can lead to higher content qualities** than would be possible under vertical separation. This is mainly because the platform internalizes the benefits of a higher quality with respect to its integrated content provider. That is, it does not 'tax' its content provider with a revenue share and thus provides the integrated content provider with an additional incentive to invest in quality, which ultimately also benefits consumers. However, the greatest consumer surplus is only achieved if the vertically integrated content provider is indeed the high-efficiency provider. In this case, vertical separation would reduce welfare for consumers. If however, the integrated provider is of low efficiency, then the effect of vertical separation on welfare is not clear and it depends on the specific competitive environment whether consumer welfare is reduced or not.

We note that the previous results were derived in the context of platforms that act both as an intermediary of content, and as a content provider itself. A very **different situation could arise in advertising supported media platforms, where the platform is vertically integrated** with intermediaries for the trade of advertising. This issue is part of the scope of ongoing competition inquiries mentioned above.

Existing theoretical literature on vertically integrated digital platforms where the platform operator also offers its own products or services on the platform indicates that vertical separation and **vertical LOBRs seem to be an option when the main dimension of competition on the platform is the *price* rather than the *quality*** of products and services. While this may apply to e-commerce markets, it does not seem to apply to search markets. The use of LOBRs in other kinds of vertical integration scenarios would require additional analysis.

However, concerning e-commerce markets, one should also be careful with vertical separation. Hagiu, I and Wright (2020)[308] argue that a vertically integrated e-commerce platform can also have a positive effect for consumers because it exerts downstream competition on innovative sellers (not the least because innovative products may be copied by the integrated platform operator), which results in lower prices for consumers; this effect can be stronger than the negative effect from less competition due to this competition. This theoretical finding seems to contradict the empirical findings of Zhu and Liu (2018) and Wen and Zhu (2019), however, who find that prices of independent sellers tend to increase after the platform enters or threatens to enter their product market (see Section 4.1.4). Moreover, Hagiu et al (2020) suggest that other policy measures, in particular a non-discrimination obligation that prevents the e-commerce platform operator from biased intermediation, would be more effective than vertical separation. In practice, it is difficult to see, however, how such non-discrimination remedies can indeed by monitored or enforced, even if this would be done by a specialised authority (see Feasey & Krämer, 2019[309] for a comprehensive discussion of this issue).

Finally, Krämer and Schnurr (2018)[310] highlight based on a literature review that **even a non-integrated platform may have an incentive to bias intermediation** (i.e., to steer consumers

[307] Krämer, J. and Zierke, O. (2020). Paying for prominence: The effect of sponsored rankings on the incentives to invest in the quality of free content on dominant online platforms. Working Paper. Available at https://ssrn.com/abstract=3584371
[308] Hagiu, A., Teh, T. H., & Wright, J. (2020). Should Amazon be allowed to sell on its own marketplace? Available at: https://ap4.fas.nus.edu.sg/fass/ecsjkdw/hagiu_teh_wright_may2020.pdf
[309] Feasey, R., & Krämer, J. (2019). Implementing effective remedies for anti-competitive intermediation bias on vertically integrated platforms. Centre on Regulation in Europe (CERRE). Available at: https://www.cerre.eu/publications/implementing-effective-remedies-anti-competitive-intermediation-bias-vertically
[310] Krämer, J., & Schnurr, D. (2018). Is there a need for platform neutrality regulation in the EU?. Telecommunications Policy, 42(7), 514-529.

away from the best 'match' or 'offer') when it is allowed to sell prominence on the platform. Interestingly, the 'congruent' and 'conflicting' payoffs environment is decisive here as well. Biased intermediation is only likely under a 'conflicting payoffs' environment, i.e., when firms on the platform compete on prices. Thus, under a 'conflicting payoffs' environment, vertical separation may not even be enough to alleviate the platform's incentive problems altogether.

5.1.4.3 LBORs for ancillary data services

In our view, in data-driven markets **policy makers should consider LOBRs first and foremost in the context of *ancillary services*** (see Section 4.1.3) These allow digital incumbents to potentially collect user data from a large range of third-parties, which diminishes the user data that these third-parties have exclusive access to, and in turn, reduces the ability of third-parties to develop a data-driven competitive advantage of their own. While incumbents receive user data in this way from various markets, they do so without competing in the market directly, and therefore without innovating in it. Thus, they do not make use of their data-driven network effects and their lower marginal costs of innovation in these markets, which may provide an overriding efficiency rationale to what has been discussed in the context of data siloing more generally.

In particular, we believe policy makers should look into ancillary **data services of already data-rich firms with 'strategic market status'** that present

- **identity management** (single sign-on) services,
- **payment services**, or
- **other services that can track user behaviour** across a wide range of third-party websites (e.g. web analytics services that third-parties embed in their website code).

And carefully review whether these could be (horizontally) separated and provided by a structurally independent firm, dedicated to that service. In these cases, structural separation seems workable, and, due to the nature of ancillary services, the boundaries of the services can be relatively clearly delineated. Thus, in these cases, the issue of 'unscrambling the scrambled eggs' is less pronounced.

Requiring data-rich incumbents to divest these services could also allow entry and growth of existing independent firms that could fulfil these important functions in the data ecosystem **as more neutral intermediaries**, i.e. as intermediaries that do not face incentive problems because they operate at different levels of the value chain. There could also be several of these firms and competition, but it is likely that here network effects and concentration tendencies will also settle in eventually. However, this would not be as problematic as in a case of a digital conglomerate firm, because the commercial activities that these firms should be allowed to pursue should be limited by LBORs as well to maintain a healthy balance of user data agglomeration in the data ecosystem. These firms could then offer these services independently to other digital market players in the same interoperable way as third-parties can use the services today, e.g., by means of protocols and interfaces.

Unbundling of these ancillary services would not even have to mean that the services would need to be offered for a price. Firms can still pay with their data if the ancillary service provider is allowed to monetize this user data by selling it in an aggregated format. In this case, it needs to be made sure, however, that the collected user data is made available to a wide range of firms, e.g., based on fair, reasonable and non-discriminatory (FRAND) prices. In this way, the data-driven network effects of the collected data could be maintained and shared, instead of having them only confined to a very small set of firms. Thus, there is reason to believe that in this case structural separation and unbundling would increase efficiency (both in the short and long term) rather than undermine it, as would be often the case in many of the other scenarios that we have discussed.

Unbundling of ancillary data services would also be welcomed from a privacy perspective, because it would limit the ability of the already largest and data-richest digital firms to extend their data collection efforts to almost every conceivable activity of users online, especially in those markets where they do not have an own consumer-facing service and therefore not even an immediate customer relationship with a given user.

5.1.5 *Privacy Enhancing Technologies*

Finally, a myriad of **technical solutions that aim at 'privacy by design'** can play an important role in limiting excessive data collection and data agglomeration. We use the umbrella term privacy-enhancing techniques (PETs) to refer to these technical solutions. Next to the possible behavioural remedies discussed above (and in Section 5.2), such technical remedies can play an important role in practice, either as a complementary or as a stand-alone remedy to level the playing field between data-rich incumbents and (potential) competitors. In any case, this option should not be neglected in the policy debate on the regulation of access to data.

PETs are subject to a considerable amount of past and ongoing research, and it is not possible to discuss them here in any detail. In general, PETs can be differentiated into two categories. Firstly, ***soft* privacy technologies** that assume a central trusted third-party exists which can undertake the processing of data. The focus of these approaches is to establish a secure communications channel to the trusted third-party (e.g., SSL encryption), to control access to data at the third-party (e.g. Oauth/Open Authorization), and to make sure that any inferred data that is passed on by the third-party is aggregated in such a way that it preserves individual privacy and cannot be de-anonymised (e.g., k-anonymity or differential privacy; see also Section 5.2.3.1).

Secondly, ***hard* privacy technologies** assume that no third-party can be trusted. Often these techniques build on decentralised data processing (e.g. data processing on the local device) and data minimisation, but they would also include technologies that enable decentralised and secure storage of data, such as blockchain and distributed ledger technologies.

The use of PETs as a technological remedy in digital markets will require an intimate technological knowledge of the authority that administers it and is always highly context-specific and targeted at a particular issue. There are certainly **no one-size-fits-all solutions** and no general recommendations on particularly promising PETs can be made, as each approach comes with several trade-offs that need to be carefully evaluated in each context. On the downside of PETs are often issues like scalability and performance (especially in the context of hard PETs), and lack of convenience for users. Moreover, the use of PETs will usually **require significant technical and institutional changes** to the status quo and are usually not interoperable with existing installations. Thus, there is likely a need to upgrade and standardize client-side and server-side (software and hardware) installations, possibly for a large heterogeneous installed base. Associated with this is a chicken-and-egg problem where one side (say the server side) requires the other side (say the client side) to upgrade first before it makes sense to follow suit and the other way around. In any case, implementing such systemic changes as part of a remedy would possibly require far-reaching powers of the administering authority to coordinate required actions, and deep industry and technical knowledge to perform this task well.

Consider the context of **digital advertising** for example.[311] Here, many PET solutions have been developed and are being developed that can be implemented in local devices, specifically browsers. For instance, *Google* has put forward a proposal known as *Federated Learning of Cohorts (FloC)*[312] which is part of Chromium's Privacy Sandbox. The main idea is that instead of observing the browsing behaviour of individuals, the behaviour of cohorts ("flocks") of similar users are observed. The main intuition behind Federated Learning, in turn, is to conduct machine learning in a decentralised and local manner on the device. Using local (deep) data on user behaviour, partial model updates are trained directly on the device. The partial updates from different devices are then sent (in a secure and possibly aggregated way) to a central server, where the partial updates are integrated into the full model. The advantage is that (deep) user data never leaves the device, and only derived data models are shared. But FloC does share information on which "flock" or cohort a user belongs to, and in combination with other techniques, such as browser fingerprinting a user's identity may nevertheless be compromised. Nevertheless, such an approach would to some degree enable targeted advertising while significantly limiting the collection of personally identifiable data compared to current solutions, e.g., based on cookies. The *privacy-focused browser Brave* has proposed a different solution, where the targeting of ads is also pursued locally. In this case, an ad server would

[311] The following paragraphs borrow from the comprehensive discussion of this issue in Appendix L of the CMA's (2019) "Digital Platforms and online advertising" interim report , available at: https://assets.publishing.service.gov.uk/media/5df9efa2ed915d093f742872/Appendix_L_Potential_approaches_to_improving_personal_data_mobility_FINAL.pdf

[312] https://github.com/jkarlin/floc

push both a catalogue of possible ads, and a targeting model to the browser, and the browser decides locally which of the ads in the catalogue is displayed. With this solution, no user data would need to leave the device.

The preceding example shows that, in specific applications, PETs can be used to limit the centralized collection and aggregation of user data while achieving comparable outcomes as if such user data were shared. While decentralized PETs like Federated Learning may be a promising approach for targeted advertising, this does not seem to be a plausible solution for other use cases such as **'search'**, however. In a meaningful and performant application of search, it is required that search queries are revealed to a central instance, which responds with a results list. Pushing the web index and the search algorithm to the device, such that search can be performed locally in a privacy-preserving way is not an option. Likewise, the search is inherently individual and cannot be done in a federated way.

5.1.6 Summary of possible remedies that limit data collection

In this section, we have considered several potential data access remedies that aim a limiting the ability of data-rich incumbents to collect ever more user data. The general problem with these sets of remedies is that they seek to achieve a more **level playing field in the digital economy by breaking the data-driven network effects of the incumbents**. This is associated with diminishing the efficiency of the incumbent and also the ability to create value from data more generally. Many remedies in this category would **severely diminish economies of scale and scope of data, and ignore the non-rival nature of data**, which makes it possible to share data that is collected by one firm with many others. Data minimisation is, of course, a value in its own right from a privacy perspective, but our assessment here is mainly based on economic rationales in view of facilitating market contestability and niche entry.

For these (economic!) reasons, we are particularly sceptical about remedies that would involve **data siloing or shorter data retention** periods. In the latter case, an additional concern is that this would rather benefit incumbents (which have a continuous inflow of large amounts of user data) and be to the detriment of entrants.

Furthermore, we are also sceptical about banning incumbents from **buying into default settings**, such as being the default search engine on a browser. Here the main reasons are twofold. Firstly, this would likely create short- to mid-term inefficiencies in cases where data-driven network effects are indeed key for the quality of a service (like in search, because consumers would then deliberately be steered to an inferior digital service. Secondly, buying into defaults can be seen as a 'tax' on the incumbent which redistributes some of the value created from data to other digital market participants that have strong consumer-facing products (like browsers or devices). Prohibiting incumbents of buying into defaults would therefore severely diminish the possibility to redistribute wealth in the digital ecosystem and destroy the business model of providers of independent browsers.

Likewise, **Line of Business Restrictions (LOBRs)** should be considered as policy measures of last resort. Especially in data-driven platform markets, they have associated with several potentially severe efficiency losses that may overcast the benefits, namely enabling niche entry in horizontal markets and resolving the incentive issues in vertical relationships. After reviewing the theoretical literature on a vertically integrated platform where the platform acts both as an intermediary and a competitor on the platform, we conclude that, if at all, vertical separation or vertical LBORs may be an option to alleviate incentive problems in such cases where the main dimension of competition on the platform is 'price' rather than 'quality' of products and services.

We do, however, also identify a potential area of intervention with LBORs that policy makers should look into. Limiting business activities of incumbent digital conglomerates in specific key (data) services that enable them to establish a global reach on economic and social activities across the web, especially **financial intermediation services and identity management services**, should be explored in more detail. We believe that in this case, the efficiency losses are less likely to be outweighed by the potential benefits from competition, let alone privacy.

Finally, policy makers should also consider the possibility of technical remedies in the form of **Privacy Enhancing Technologies (PETs)** alongside potential behavioural remedies. However, such approaches can only be used in very specific contexts and are likely to require significant coordination and implementation efforts. There are also concerns with respect to legal certainty and whether this

is indeed the right regulatory tool (Tinbergen principle) to address economic policy goals. Moreover, imposing PETs as a technical remedy will likely require a skill set that is not typically present in regulatory authorities. While PETs are generally beneficial for users' privacy, specific proposals for PETs should also be scrutinised closely as to whether they would indeed enable entry by new firms, or rather increase entry barriers by further advantaging the user data collection capabilities of incumbents relative to entrants.

5.2 Remedies that facilitate access to 'broad' raw user data through bulk sharing

5.2.1 General comments on mandated data sharing

The unique characteristic of data as a bottleneck resource, as opposed to material bottleneck resources, is its non-rivalrous nature. Thus, the bottleneck can in principle be resolved by enabling non-exclusive access to it. Furthermore, a special characteristic of data is that is has a low specificity and can be repurposed and re-used to create more and add value. **Both of these characteristics strongly point to 'data access' or 'data sharing' as ideal remedies** to resolve potential data bottlenecks and barriers to entry created by data.[313] These sets of remedies are also generally more desirable from an efficiency point of view because they are aimed at increasing the efficiency of third-parties, rather than limiting the efficiency of the incumbent.

There are, however, also caveats and limits to data sharing as a remedy in a competition context, especially in the context of user data. **Trade-offs occur** particularly due to privacy concerns and conflicts of laws concerning privacy regulation. But also economic trade-offs occur because data sharing can not only increase the potential to create value (through re-purposing and innovation), but also diminish the incentives to collect data in the first place, which would then deprive the potentials of value creation from data. There is, hence, a broad consensus on the fact that data access and data sharing remedies in the digital economy, if they are pursued at all, should **focus on raw user input data (volunteered and observed data)**. Such raw user data is provided by users effortlessly and 'en passant' while using a service and can be recorded automatically and therefore at virtually zero marginal costs by the service provider (see also Section 3.1). Inferred and derived data about users, however, should usually be considered off limits, especially for ex-ante regulation. Such data is the result of innovation efforts (e.g. in data analytics) with the intent to derive actionable insights that are then the basis for competition in digital markets. Raw data, by contrast, does not provide actionable insights per se and therefore has zero economic value until it is processed and analysed.

Two additional economic arguments generally justify data access and sharing remedies for data incumbents that are focused on raw user data. First, Acemoglu et al. (2019)[314] and Bergemann, Bonnati and Gan (2020)[315] point to a **particular externality of user data, which they call 'social data'**. That is, a data incumbent that already has collected a very large sample of user data can use this data to make predictions also about new users that are outside of the sample. In other words, data collected about an individual user is not only informative about that specific user but also about similar users, such as in the creation of virtual 'twins' to target advertising. This implies that for data incumbents, who already have access to a large trove of raw user data, the marginal value of obtaining more user data approaches zero.[316] Consequently, exclusive private control over user data implies an inherent market failure, because the agglomeration and exclusive use does not maximize the benefits from this data.
Second, Martens (2020)[317] highlights that **data is not a homogenous good** in the sense that data can be provided at different levels of detail (e.g., 'depth' and 'breadth' in our terminology). Thus data sharing does not necessarily mean that the original data controller is not able to retain a competitive advantage. Indeed, in the context of personal user data, privacy regulation usually presents a natural limit to the detail of data that can be shared. As we have argued in Section 4.2.2, it is therefore not realistic to assume that data sharing and data access remedies would enable contestability of the

[313] For the following discussion the direction in which data is acquired, i.e., whether it is pushed to the third-party ('shared') or pulled by the third-party ('accessed') will often not matter. Therefore, we will largely use 'data access' and 'data sharing' as synonymous in the following in order to ease understanding.
[314] Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). Too much data: Prices and inefficiencies in data markets. Centre for Economic Policy Research Discussion Paper DP14225. Available at: https://repec.cepr.org/repec/cpr/ceprdp/DP14225.pdf
[315] Bergemann, D., Bonatti, A., & Gan, T. (2020). The economics of social data. Available at: https://arxiv.org/pdf/2004.03107.pdf
[316] Note that this argument relates mostly to the 'breadth' dimension of user data.
[317] Martens, B. (2020). Data access, consumer interests and social welfare An economic perspective on data

incumbent in the narrow sense, because the detail of raw user data that can be shared is inherently limited in practice.

The case studies have shown that in **many applications third-parties need raw user data with depth *and* breadth to develop a competitive data service**, which may then allow them to develop data-driven network effects on their own. Particularly valuable are behavioural user profiles that allow tracing an individual across several choices. While traceability is important, personal identifiability is often not necessary. For example, recommendation systems benefit from deep user profiles when searching for similar users in the data set, but this does not require to personally identify such users (see Section 3.3 for a similar observation with respect to search). However, sometimes traceability is difficult to achieve without identifiability. Moreover, identifiability can often contribute additional value or is required as a necessary input in a specific step of the value creation process, such as for the re-identification of users in the case of targeted advertising (see Section 2.3).

In the following, we, therefore, suggest using two types of data access and sharing remedies in concert. (i) The first type of remedy is to facilitate **access to broad user raw data** for third-parties. This can only be achieved by bulk sharing of sufficiently anonymised raw data. Such data will therefore generally lack depth but is in terms of breadth representative of the raw user data that the original data controller has access to. (ii) The second type of remedy, discussed in Section 5.3, is to facilitate **access to deep user data**. This data contains personally identifiable information or at least allows traceability of an individual. Such data cannot be shared in bulk but requires the consent of each data subject anytime it is shared with a third-party. Thus, the sum of data that is shared in this way generally lacks breadth, because it is unlikely that a sufficiently representative sample of users will consent to data sharing for a given third-party. Taken together, both types of remedies would allow, in the best possible way, to share broad and deep user data with third-parties. In the remainder of this section, we discuss the details and main trade-offs involved for each type of remedy.

5.2.2 Scope of access for mandated sharing of broad user data

Next, we discuss the scope and detail of access if mandated sharing of broad user data with third-parties should be imposed as an ex-ante remedy. Note that we do not (yet) discuss here which firms should be subjected to providing access. This issue of governance is briefly discussed in Section 6.2 and considered thoroughly in the companion report by Feasey and de Streel (2020). In the following, we derive several principles for broad user data sharing which are motivated by our preceding analysis and which are meant to trade off the potential gains for competition and innovation with the potential risks to users' privacy and the maintenance of incentives for developing innovative services and deriving actionable insights from data.

Principles for the scope of mandated sharing of broad user data with third-parties

1. **Only raw user data** (observed and volunteered) may have to be shared; but not derived insights from such data
2. Only data that was created as a ***by-product* of consumers' usage** of a dominant service may have to be shared (e.g., search queries, likes, clicks, or location); but not (volunteered) user data that represents the essence of the service itself (e.g., documents uploaded to a cloud storage provider, posts on a social media site, customer reviews on a reviews' site, or GPS data from a geo-tracking app)[318]
3. Any bulk data sharing must be done securely **and be anonymised** sufficiently to preserve individual users' privacy
4. Generally mandated data sharing should be done in **real-time and continuously**, making use of appropriate technical interfaces (APIs) and standards.

While these principles should be seen as a general guideline for any mandated bulk sharing or user data with third-parties, it is evident that the precise scope and terms of access need to be carefully **assessed on a case-by-case basis** by the competent authority. This will be feasible from an administrative point of view because we envisage that only a small subset of firms will be subject to such bulk data sharing (see Section 5.2). We attempt to discuss some of the specific trade-offs that would arise if bulk sharing of user data were mandated in the context of search, e-commerce and media markets (i.e., the case studies review in Section 0) in Section 5.2.4.

Two general trade-offs are worth discussing more generally. First, we emphasize in the second point of our guidelines that data sharing should focus on user data that was created as a by-product and not as the main product of a given service. The lines between the two may become blurred. For instance, we stipulated in our examples that observed GPS data of users should not be shared if the dominant service in question were an app whose main purpose is to record GPS data, e.g. so that it can be analysed and processed in a mapping software later. This is to make sure that data sharing does **not undermine the main value proposition and business model** that a provider pursues. For example, the respective provider's service may be dominant in the market because it has found a way to pinpoint the location of a user with higher accuracy than any other service. Contrast this with a situation in which the service is a maps software, whose main purpose and value proposition to end-users (!) is to help them to navigate to a certain location. This app needs to collect location data as well, but one might argue that this time location data is collected as a by-product of navigation, and the main purpose of the app (to the end-user at least – who's perspective is what matters for this analysis) is not to collect location data about that user. Similarly, customer reviews on restaurants in a service like 'Uber Eats', whose main purpose is to allow customers to order food directly through the service, may be viewed differently from customer reviews on a service like 'Yelp', whose main purpose is to collect and provide those reviews (also to third-parties), but where it is not possible to order food directly. It may sometimes be difficult to walk this line. In particular**, it should also be taken into account whether viable commercial offers for bulk data access already exist**. If this is so, mandated sharing may not be necessary. Focussing on regulated **sharing of data that has been created as a by-product also has the advantage that it will often be justified to share this data for zero** without undue harm to the business model and innovation incentives of the data provider. Nevertheless, we see a need to determine the appropriate conditions for sharing on a case-by-case basis (for more details see Section 5.2.4). If, however, data ought to be shared that is not a mere by-product, then the appropriate price for this data will usually be non-zero. This, in turn, gives rise to numerous questions and complexities concerning the determination of a regulated price that is fair reasonable and non-discriminatory (FRAND). This is discussed in more detail in the companion report by de Streel and Feasey (2020).

A second main trade-off arises by the need to **balance users' privacy** with maintaining enough level of detail in the data that it is valuable for third-parties under a broad set of possible applications and allows them to derive novel insights through data analytics and machine learning. Some observers seem to question that it will ever be possible to balance this trade-off, as methods and tools for de-anonymisation are continuously being improved and even relatively little detail may

[318] An individual user may port this data, on an individual basis, to another provider, however. We discuss this in Section 5.3

already reveal a person's identity (see, e.g., Rocher, Hendrickx and de Montjoye 2019)[319]. However, we argue that it is generally feasible to balance this trade-off and to share user data in a meaningful and privacy-preserving way using technical *and* institutional means. We briefly discuss this in the following subsection. We already note here that the appropriate choice and combination of means to preserve privacy will again be highly context-specific and may also change over time as more or fewer data points are being made available as more or fewer firms are being mandated to share data. The level of detail may also depend on who the data is being provided to and on what terms and conditions this occurs. Nevertheless, anonymised user data will never have the same 'depth' as the original data set due to this trade-off, of course. As indicated above, this is one of the reasons why it is unlikely that data sharing enables a third-party to contest (in the narrow sense) the incumbent who had provided this data; although data sharing can be a stepping stone for innovation and growth that may eventually lead to contestability. Sharing of broad but less deep user data, which is representative of the whole user base, is useful in many applications and would certainly improve data availability to third-parties (including research) over the status quo.

5.2.3 Technical and institutional means to preserve privacy in shared data sets

5.2.3.1 Technical means: Anonymisation

The risk of de-anonymisation in a particular data set depends crucially on the uniqueness of the attributes associated with different individuals. It is therefore generally not enough to just remove a personal identifier (e.g., the combination of full name, birthday and place of birth) and to replace it with a pseudo-identifier (e.g., a unique combination of numbers and letters). Although it might not be immediately obvious anymore who is associated with a given data record, the values of the remaining attributes in the data set (e.g., the combination of blood type, zip code and age) may still uniquely identify an individual. This is the more likely, the more unique individual values are (e.g., a very rare blood type or a very high age). '**Anonymity' is therefore not a discrete zero-one concept but rather a statistical concept** that relates to a particular probability that an individual may be re-identified.

In computer science, two concepts are frequently used to describe the degree of anonymity in a given data set. The first concept is ***k-anonymity****:* A data set is said to have k-anonymity if the information for each person contained in the data set cannot be distinguished from at least k-1 other persons who are also contained in the same data set. Consequently, the larger k, the larger is the degree of anonymisation of a data set. K-anonymity can generally be achieved by suppression of attributes (e.g., deleting name, dates or address) or by a generalisation of attributes (e.g., transforming names to initials, dates to years, and addresses to zip codes). K-anonymity usually does not involve any randomization of attributes and it can be shown that in large data sets, especially 'deep' data sets with many attributes, anonymity may nevertheless be compromised.

A second, more recent and more sophisticated concept is ***differential privacy****.* Describing the concept would go far beyond the purposes of this report. Roughly speaking differential privacy is not a discrete concept (as k-anonymity), but a probabilistic concept and requires randomization of attribute values (e.g., adding some random noise to GPS data). The goal is to create a data set for which it is not possible (with some statistical guarantees) to know whether an individual's data is contained in the data set. This is important because de-anonymisation attacks typically match data from different data sets from which it is known that they contain a given individual. This can be achieved, for example, by running several similar queries to a database, to obtain (anonymised) data sets that differ only by the entry of one person. While data sets with a k-anonymity property are susceptible to such attacks, data sets with differential privacy are not, due to randomization. There are several algorithms to achieve differential privacy, and this is subject to ongoing research in cryptography. In practice, it may be difficult and computationally burdensome to achieve differential privacy, especially if data is shared continuously. A more practical approach is therefore not to store accurate data about individuals at all, but to add some noise already when data is collected. This is a technique that is already applied by Apple and Google for select applications in iOS/macOS and Chrome.[320] This also highlights that differential privacy is not just a theoretical option, but can indeed

[319] Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9. Available at: https://www.nature.com/articles/s41467-019-10933-3

[320] Green, M. (2016). What is differential privacy? https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/

be applied in the context of large-scale data collection as is typical for prominent digital services. This may also mean, however, that regulated firms may not only be mandated to share their data but also mandated to collect (or rather *not* collect) their data in a certain way, to enable privacy-preserving sharing of that data later.

5.2.3.2 Institutional means: Data trusts and data sandboxing

Next to such technical means, there are also institutional means to protect privacy, which can also be combined. A common institutional proposal is to establish ***a trusted data intermediary (data trust)*.** To ensure this, the trust needs to be independent of the regulated entity, of course. The main idea is that user data (from the various entities that are mandated to share data) is collected by a data trust in its original raw and detailed form (see, e.g. Prüfer 2020). The trust could then combine the data and anonymise it properly. Such anonymisation of the joint data set directly would be preferred over anonymisation of separate data sets at the source because it would reduce the risk of de-anonymisation through re-matching of the different data sets, each of which may have different attributes omitted or generalised.

Moreover, the data trust may not need to reveal any raw data directly but **could act as a *data sandbox*** instead. This means that third-parties would need to submit their algorithm for analysing the data to the trust, who would then run it on their behalf on the detailed raw data. The third-party would receive back the trained algorithm, but never see the raw data itself. Data sandboxing could also be applied to the original data source directly (Prüfer, 2020).

However, there are several practical issues with data trusts and data sandboxes, especially when applied in the context of the digital economy. First, for all practical purposes, a data trust would **require an enormous infrastructure** to be able to store, aggregate and anonymise the data (continuously) in any meaningful way. For example, Google Search alone processes over 80,000 search queries every second on average, which translates to almost 7 billion searches per day.[321] It seems one would have to duplicate the data centre infrastructure of Google and Amazon to achieve this. Who would then finance and operate this, and be liable in case of failure or data breaches?

Likewise, data sandboxing is an intriguing theoretical idea, but it would require an even larger infrastructure to have enough computing power required for running probably complex algorithms on the data. Since these would operate on the detailed raw data, it would also require enormous effort and expertise to make sure that the algorithms do not compromise privacy. If algorithms are run directly on the infrastructure and raw data of the original data controller, then this would also put a significant computational burden and cost on the regulated firm. In turn, this would require some compensation and quite possibly give rise to issues of margin squeeze or sabotage (Mandy and Sappington, 2007[322]), each of which would raise the need for additional regulation. It also needs to be feared that the original data controller would be able to acquire business-sensitive information about the third-parties through the algorithms that are run on its infrastructure.

Nevertheless, we entertain the idea that a **data trust and data sandboxing (at a data trust) may be feasible if confined to subsets of the data, particularly with a focus on recency, and if confined to a few select algorithms that may be trained at any given time**. Indeed, the EuroHPC[323], which is a 1 billion Euro joint initiative between the EU and European countries to develop a world class supercomputing ecosystem, may present a capable technical infrastructure for such a European data trust. The data trust should not be the only source for access to broad user data, however, especially because it can only offer limited access. The original data controllers should also make broad data available through APIs, albeit data this would need to be anonymised to a larger degree.

5.2.3.3 Unlawfulness of de-anonymisation

Finally, for completeness, we mention that policy makers could also use **legal means to disincentivize third-parties from attempting to de-anonymise data on purpose**. As described above, de-anonymisation of deliberately anonymised data sets usually requires effort and intent on the side of the data acquirer. In the applications that we have in mind here, it is not conceivable that

[321] Internet Live Stats, 2020, https://www.internetlivestats.com/one-second/#google-band
[322] Mandy, D. M., & Sappington, D. E. (2007). Incentives for sabotage in vertically related industries. Journal of regulatory economics, 31(3), 235-260.
[323] https://eurohpc-ju.europa.eu

de-anonymisation would just happen by accident. So what if deliberate attempts to de-anonymise shared data sets would be illegal under European law (Prüfer 2020)? At first, this seems to be something that is pure rhetoric and difficult to enforce and monitor. However, at second thought the same holds for cartels, and in this context, whistleblowing has proven to be very effective to obtain insiders' knowledge. There is now a system in place where individuals (be it from inside or outside the firm) can inform the European Commission about illegal actions anonymously, and where the first firm to unveil an illegal action can apply for leniency and avoid any fines.[324] A similar system could also be used in the context of de-anonymisation.

5.2.4 The devil is in the detail: Possible broad data sharing remedies in the case studies

In the following, we attempt to make some progress on the debate on how specific data sharing remedies could look like in the case studies that we covered. It is clear that we can only discuss the tip of the iceberg here and that significantly more thought would need to be put into these truly hard problems. Nevertheless, we believe it is useful for the progression of the debate to detail some of the more specific trade-offs that occur in each case study. It will also be seen that in different contexts, similar trade-offs occur and that solving one of them is informative for the other.

5.2.4.1 Search: Query logs

In general search, the **data bottleneck lies in the search queries, associated context information and behavioural data** on how users interacted with the search results (see Section 2.1). The data bottleneck is not the web index per se, however, as this data can be duplicated by (potential) competitors!

Any shared data must therefore at least contain information about the search queries that users have presented to the search engine provider. This already brings about one central difficulty. **Search queries are inherently personal and can reveal significant information about an individual**. This may also reveal a person's identity relatively easily. A famous example is the case of Ms. Arnold, who was identified from a list containing 20 million web search queries conducted by a total of 657.000 Americans over a period of just three months. Although the data set was released by AOL in a pseudo-anonymised way (evidently not respecting k-anonymity or differential privacy), she was re-identified based on her search queries alone.[325]

Since web queries are based on text, it is not as straightforward to add some noise to the search terms without rendering them useless. Moreover, as has been pointed out in the case study in Section 2.1, it is precisely the rare search terms that are particularly valuable for training a prediction model improving ranking quality.

The risk of re-identification is less pronounced if one would not associate specific search queries with a unique user identifier, which allows associating different searches of an individual over some time. However, without such a user identifier, the data set loses traceability, which is an important property for many applications. The more data on an individual user is released and associated with certain queries, the easier will it be to re-identify that user. At the same time, and inevitably so, the data set becomes more valuable for repurposing. For example, if information about a user's age, gender and location is released with her or his search queries, then that information could be used to derive information about possible medical conditions in different populations and/or locations. This general idea was used, for example, by Google Flu Trends to predict the spread of the flu based on users searches for symptoms associated with flu.[326]

The **anonymisation of search logs, while preserving useful information**, is a relatively recent and emerging field of research (Hong et al. 2009)[327], but also in this domain of computer science, progress is being made quickly. Promising developments seem to be the creation of 'synthetic search logs' which contain plausible search sequences, but are created from a machine learning model and do not relate to an actual person (see, e.g., Krishnan et al. 2020).[328]

---

[324] See https://ec.europa.eu/competition/cartels/whistleblower/index.html
[325] Barbaro, M. and Zeller, T. (2006). A Face is Exposed for AOL Searcher No. 4417749. New York Times. Available at: https://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0
[326] See https://www.google.org/flutrends/about/
[327] Hong, Y., He, X., Vaidya, J., Adam, N., & Atluri, V. (2009, November). Effective anonymisation of query logs. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1465-1468).
[328] Krishnan, U., Moffat, A., Zobel, J., & Billerbeck, B. (2020, April). Generation of Synthetic Query Auto Completion Logs. In European Conference on Information Retrieval (pp. 621-635). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-45439-5_41

A second major challenge is to **define the scope of the contextual information relating to the search results page properly**. Aggregate or even individual search query data is only one part of the relevant information that users reveal to a search engine. The other part is how they have interacted with the search results page, for example, which links were clicked after a given search and in which order? But it may sometimes be even more informative which links consumers did *not* click and thus did not find relevant, after a given search. For a proper assessment of clicks, it would also be necessary to know which other elements were shown on the search results page in addition to the organic search results. For a long time, Google's search results page does not only contain "10 blue links" anymore, but also, and depending on the search query, sponsored search results and other 'boxed' elements such as a news carousel, flight search, a shopping comparison or an immediate answer to the search query are displayed. An increasing percentage of search sessions end with the search results page, and consumers never follow up and click on a search result. It is estimated that so called Zero-Click Searches amounted to about 50% of all searches on Google.com in June 2019 (Fishkin 2019)[329]. Likewise, research has shown that clicks on organic search results are heavily influenced by whether and how sponsored search results and 'boxed' results are presented (Edelman and Lai, 2016)[330].

The search results page is already inferred data of the search engine, and forcing the release of detailed information about the search results page pertaining for every query would go too far and undermine past and future innovation efforts. However, it may be justified to release such information for samples of queries or to limit the details of the data relating to the search results page. One could release, for example, only the first clicked result.

To advance the discussion on the appropriate scope of shared search logs, **we suggest to think in three main categories: i) data on the query itself, ii) data on the search results page, and iii) data on the user**. Figure 5 exemplifies which pieces of information can belong to each category.

DATA ON THE QUERY	DATA ON THE SEARCH RESULTS PAGE (SERP)	DATA ON THE USER
Keywords (e.g., raw search string, synthetic search string)	Clicked URLs (first clicked result, last clicked result, all clicked results)	Unique identifier
Timestamp (e.g., week, day, hour, seconds)	Zero-Click search (yes/no)	Device metadata (e.g., mobile/ desktop, browser metadata)
Connected queries in the same session	Results ranking (top 3, top 5, top 10)	Location data (IP-address, GPS)
	Layout of the SERP (sponsored results, one-boxes)	Other available user attributes (e.g., age and gender from account data)

Figure 5: Categories and scope of search data to be considered for sharing

This is certainly not a complete list, but it invites policy makers to think how different data, each at various level of granularity (listed in parentheses) can be mixed and matched from the different categories, and this would result in significantly different data sets that may be shared. The preceding discussion also shows that it is probably not useful to think about 'the' data set that should be shared. Rather, it seems to fruitful to think about **two different types of mandated access to data**:

- ***Publicly shared data**:* A data set that is made publicly available through APIs which is highly anonymised (e.g., no attributes from the 'user data' category) and contains only keywords with limited contextual data (e.g., coarse timestamp, coarse location, first clicked

[329] Fishkin, R. (2019). Less than Half of Google Searches Now Result in a Click. SparkToro. Available at: https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/
[330] Edelman, B., & Lai, Z. (2016). Design of search engine services: Channel interdependence in search engine results. Journal of Marketing Research, 53(6), 881-900.

result). Google, for example, already provides some limited information of this sort in Google Trends.

- ***Individually shared data***: A more detailed data set may be made available to third-parties after a vetting procedure by the regulatory authority. The authority could verify legitimate interest to access more detailed data, and make a data set available that caters specifically to those needs. The access seeker could also be subjected to higher responsibilities and safeguards in this context, e.g., for de-anonymisation.

These two modes of access may be able to balance the inherent different trade-offs of making broad data available for purposes of innovation while protecting consumers' privacy. Furthermore, in the latter case, where more detailed data is shared individually, consumers' may additionally be given the opportunity to opt-out of such data sets.

5.2.4.2 E-Commerce: Sales data, reviews and product queries

In the context of e-commerce there are several categories of data that one may consider for mandatory data sharing.

(i) Sales data

In the context of e-commerce platforms, it is **sometimes alluded to that independent sellers would lack crucial business data on the platform**. For example, in the Flash Eurobarometer 439 Survey, 42% of the respondents said that they usually do not get the data they need about their customers from online marketplaces.[331] Note that this concerns the flow of data from the platform to the sellers and not the other way around, which was discussed in Section 4.1.4.

However, **e-commerce platforms do share relevant information with the sellers**. A seller always knows whether or not his product was sold, and typically he also knows or could know who has bought it, because the seller ships the item directly to the customer (or has tasked the platform to do so). In a study for the European Commission on platform-to-business data use, VVA (2018)[332] has found that large platforms generally offer dashboards to their business customers, where they receive aggregate performance measures. The study also finds that larger platforms usually share more and more detailed information than smaller platforms. Issues are raised, however, with respect to a firm's ability to export the data provided from the dashboards (e.g., as downloads) so that they can be transferred to other platforms.

With respect to a lack of data, business users seem to be most concerned with data about i) user identification details and ii) user behaviour data on the platform. With respect to **user identification details**, sellers seem to be most interested in the e-mail addresses of platform customers. Such information, they say, could be valuable, e.g., for promotional activities. However, the platform also has a commercial interest to withhold this information, because it may be used by the seller for circumventing the platform altogether. The platform would then merely be used for window shopping. Thus, in our view, it is understandable that such information is not shared in order to protect the investment in the platform.

With respect to **user behaviour** on the platform, sellers would be interested in clicks and browsing histories on the platform, also relating to products of other sellers. Evidently, such data could easily reveal business-sensitive data on other businesses on the platform as well, and may be used to facilitate collusion, among other things. So, again, in our view, there are good reasons not to share such data.

Taken together, **we do not see a particular issue with the access of data for sellers on the platform**. Maybe the aggregated information relating to own transactions on the platform could be improved, and export of such data could be facilitated. In this context, we currently do not see a need for additional regulation over and beyond those on transparency and fair business conducts, which are already addresses by the EU Regulation on platform-to-business relations (P2B regulation)[333].

(ii) Customer reviews

We already discussed in Section 5.2.2 that customer reviews should generally only be shared if they are derived as a by-product of the main service (here e-commerce sales), and that it can be difficult to delineate what is a by-product and what is not. Two types of customer reviews seem relevant in this context. The first type of data is **customer reviews about the seller itself**. This is a similar issue as with sales data. This data is already available to the seller, but may not be transferable to another platform. In this case, the data is by its nature even publicly available, however, so we do not see an issue mandating sharing of such data.

The second type of data is **customer reviews about certain products**. This data would usually not be associated with a particular seller, because many sellers may have the same products for sale. In any case, this data is by nature also publicly available, and we do **not see a need for data sharing regulation** here.

---

[331] https://ec.europa.eu/information_society/newsroom/image/document/2016-24/fl_439_en_16137.pdf
[332] VVA (2018). Study on data in platform-to- business relations. Available at: https://ec.europa.eu/digital-single-market/en/news/study-data-platform-business-relations
[333] https://ec.europa.eu/digital-single-market/en/business-business-trading-practices

(iii) Product queries and purchase histories

Finally, search queries and contextual search data, similar to the case of general search, are collected by e-commerce platforms. Generally, in the context of product search the same trade-offs occur as for general search, and the categories of data that we have identified there are also relevant in this context.

However, an additional aspect arises with respect to '**data on the user**' that could be shared. E-commerce platforms have first-party data not just on the clicks on products, but also on the actual purchases made on the platform. This data, especially if combined with other search data, is extremely useful for deriving recommendations. But it can also be used for various analyses without associated search data, such as for product basket analysis. Such data can be very sensitive as well, in much the same way as search query logs. Thus, if a regulator should choose to make such data available to third-parties, there must be a careful case-by-case vetting of the data access seeker, and an assessment which data exactly would be warranted for the application at hand.

In any case, obligations to share data should be reviewed even more carefully in the context of e-commerce, because there seems to be fierce competition in this domain in the near future. Specifically, e-commerce and 'product search' seems to be an area of intense competition between some of the largest holders of user data in the near future: In April 2020 Google has announced that it would waive the fees for selling on Google Shopping[334] to invite more sellers to use Google Shopping and, so it is argued, to receive a larger share of product related searches. Likewise, in May 2020 Facebook has also announced that it would launch its marketplace for Facebook and Instagram, called Facebook Shops[335], which is specifically targeted at small sellers and businesses.

5.2.4.3 Media: Data on ad campaigns and audiences

Here, we focus on advertising supported social media platforms disseminating user generated content and discuss the potential sharing of two main categories of data in this context: data on audiences and data on advertising.

(i) Data on audiences

Media platforms collect various information about their audiences, similarly as e-commerce platforms are collecting information about consumers on their platform (see Section 2.3). This includes clickstreams on with which content users interacted and how they interacted with it. Social media platforms generally have an incentive to share information about the audiences of content with the users that created it to help them to evaluate the content quality and user engagement. For this reason, **media platforms generally provide more systematic access to user behaviour than e-commerce platforms** (VVA, 2018)[336]. Some of this data (e.g., likes, comments and number of views) is even publicly available. Some may be available in limited aggregate form to the users creating content and often more so to advertisers serving ad around the content.

Similar, as in the case of e-commerce, third-parties may wish to receive more detailed data or to be able to export such data (more easily). Especially in the context of media, 'deep' data about an individual consumer's preferences seem to be more important than in other contexts, as media content can be customised and personalised in even more ways than, e.g., physical products. Likewise, as in the case of e-commerce, there may be problems with revealing detailed behavioural data about interactions with content to others than the creators of the content, such as other users or possibly influencer exchanges or multi-channel network services, as this may unduly reveal business-sensitive information about the user generating the content. **A case could be made for some sharing with other parties in the model of the kind of independent audience measurement that is accessible industry-wide for other media**, which is already happening to a limited extent with the case of YouTube's participation in the German measurement system.

There may also be a case for **sharing search logs and aggregate demand of viewers on the platform**. This may also include viewing histories (in analogy to purchase histories). The trade-offs with respect to the privacy and algorithmic learning are similar as in the case of e-commerce and therefore not repeated here (see Section 0). We do note, however, that on many social media

[334] https://blog.google/products/shopping/its-now-free-to-sell-on-google/
[335] https://www.facebook.com/business/news/announcing-facebook-shops
[336] European Commission (2018). Study on data in platform-to- business relations. Available at: https://ec.europa.eu/digital-single-market/en/news/study-data-platform-business-relations

platforms, search often plays a less important role than in e-commerce, because users are commonly presented with curated media feeds and tend to follow recommendations more often. This, in turn, gives the platform an even greater control for steering user attention. Nevertheless, in some cases, such as on Youtube or Twitch, video-specific search plays a relevant role.

Furthermore, some content providers that are present on several platforms, or have their channels to audiences, complain that some media platforms have a contractual restriction, which disallows them to (attempt to) match their audiences on the platform with that in other channels. While this issue is not necessarily specific to media platforms, it seems to be more pronounced here due to the prevalence of free content and reliance on advertising. We specifically discuss this issue therefore below in the context of data on advertising although it has also been raised in other contexts (see, e.g., VVA 2018, Section 5.2) such as app stores (Toplensky and Nicolaou, 2019[337]).

(ii) Data on advertising

A contentious issue in media platforms accrues with respect to the fair attribution of user data and associated profit opportunities in the context of advertising. Typically, independent content providers are responsible for investing in the production and quality of their content but then rely (at least in part) on a media platform to publish it and to attract a large audience for it. The media platform is also responsible for placing advertisements and monetizing the content, though sometimes content creators are also able to sell advertising inventory. The advertising revenues are often then shared between the platform and the content creator.

The process of how advertising is matched to content, and how advertising is matched to consumers is opaque (and complex) however, and subject to elaborate policy investigations (CMA, 2020). It is beyond the scope of this report to address these issues in further detail than is covered in our analysis in Section 2.3. Instead, we focus on the question of whether content creators should have access to more data about the advertisements that are associated with their content and the interaction of individual users associated with those advertisements.[338]

Indeed, **most media platforms restrict the data that content creators and advertisers can access related to advertising on the platform to certain categories of aggregate data**. For example, Facebook gives advertisers access to aggregate ad performance data through its Ads Manager. It notes in its policies that advertisers may not "use Facebook advertising data for any purpose (including retargeting, commingling data across multiple advertisers' campaigns, or allowing piggybacking or redirecting with tags), except on an aggregate and anonymous basis (unless authorized by Facebook) and only to assess the performance and effectiveness of your Facebook advertising campaigns." Neither does it allow advertisers to "use Facebook advertising data, including the targeting criteria for your ad, to build, append to, edit, influence, or augment user profiles, including profiles associated with any mobile device identifier or another unique identifier that identifies any particular user, browser, computer or device."[339] Content creators, such as page owners have access to aggregate performance data for their pages, but usually do not have access to the campaign data on any advertising that might have appeared around it, other than their revenue share from such advertising. While it is convenient to use this example, because it is very explicit on the matter, Facebook is by no means special in applying such a policy. Google Ads offers similar functionality for advertisers but there is not equivalent access to data for the content creators on YouTube around whose content the advertisement has been placed. Twitch is even more restrictive in terms of the amount of data that can be collected from user behaviour and shared.
Platforms argue that such policies are necessary to protect consumers' privacy. Whether content creators, who are the ones that have invested in the content they contribute to social media platforms, are treated fairly by being prevented from using data generated on the platform for conducting business outside of the platform is a difficult decision to make. Platforms also have invested in the infrastructure for providing media, both hardware as well as software, e.g., to curate content. They generate revenues predominantly from advertising have a legitimate business interest to protect their business model. **Whether or not a specific revenue share is appropriate for**

[337] Toplensky, R. and Nicolaou, A. (2019). Spotify files EU antitrust complaint against Apple. Financial Times. Available at: https://www.ft.com/content/73e0d448-4577-11e9-a965-23d669740bfb
[338] On a related note, advertisers also complain that they lack access to user data, and in relation to which content their advertisements were shown. See, e.g., https://www.cpomagazine.com/data-privacy/google-will-restrict-sharing-of-user-data-for-google-ads-under-eu-privacy-pressure/
[339] See Section 12 (Data Use Restrictions) at https://www.facebook.com/policies/ads/

the dissemination service that platforms offer or abuse of market power, or whether there are problems with access or fair trading, are issues that require a thorough case-by-case investigation. The outcome of the Apple vs. Spotify competition case[340], which centres on this very question, will be very informative on this matter and any potential need for intervention.

Some **middle ground may be achieved, if consumers were allowed to opt-in into the sharing of their usage data, including a unique identifier, when they consume content on the platform**. This would certainly alleviate most of the privacy concerns that the platforms have raised and allow content creators to pledge their loyal audience to trust them with their data. Such a policy may also be coupled with a measure to allow consumers to decide more freely on their own, with whom they want to share their data, and to enable them to share that data continuously trusted content creators. We discuss such approaches in the next subsection in detail.

5.3 Remedies that facilitate access to 'deep' raw user data through continuous data portability

The **current EU legal framework already contains several rules imposing the portability and the sharing of personal and non-personal data**. The most relevant of these in the context of digital markets are the General Data Protection Regulation, particularly Article 20 ("Right to data portability"), and the Digital Content Directive, particularly Article 16 ("Obligations of the trader in the event of termination"). The former applies only to personal data and can be exercised by the data subject at any time when having a contractual relationship with a service provider, while the latter applies only to (remaining) non-personal data and can only be exercised when a contractual relationship is terminated.

In a related CERRE report, Krämer, Senellart, and de Streel (2020)[341] study in detail whether this legal framework is sufficient or would need to be complemented to make personal data portability more effective in the context of digital markets. Although they do so with an emphasis on enabling consumer empowerment and innovation by third-parties, in this report, we have highlighted that **access to 'deep' personal data is also valuable and welcomed from a competition perspective, especially to facilitate niche entry**. This is especially so, because the sharing of deep, personally identifiable data, which comes with several limits that we will review below, would complement the sharing of broad user data, which we have discussed above. Building on the report by Krämer et al (2020), in the following we summarize the main issues identified with the current legal framework and propose a set of remedies and policy measures to facilitate sharing of deep raw user data beyond the status quo.

5.3.1 Limits of the status quo of data portability

Firstly, numerous technical difficulties **arise from the fact that different standards and data formats** can be used following a data portability request. In particular, the sending provider must not adhere to a certain standard and can change it at any given point in time. These uncertainties regarding standards and their perseverance can make it very costly for a new provider to offer an interface to import data. In return, this means that **more stringent and common standards for data portability** are key to ensuring that data is more widely imported and used. The provisions in GDPR, which merely call for a "structured, commonly used and machine-readable format" are not enough. If the same type of data (e.g., photos, videos, search logs) would be made available in the same format, irrespective of the provider, then it would be more feasible to develop and provide respective import adapters. More widespread availability of such adaptors and re-usability of ported data would also raise awareness among users and encourage them to port their data. The transfer could further be **facilitated by Personal Management Information Systems (PIMS)**, who could perform schema mappings between various services.

Secondly, Article 20 GDPR may not be sufficient to truly empower users in digital markets and foster competition and innovation. Often consumers want to try out a new service provider immediately, and that provider may be in need to a cold start with the users' data to offer an immediately appealing service. But the GDPR does **not give the consumers the right to immediate and very frequent**

[340] Toplensky, R. (2019). Brussels poised to probe Apple over Spotify's fees complaint. Available at: https://www.ft.com/content/1cc16026-6da7-11e9-80c7-60ee53e6681d

[341] Krämer, J., Senellart, P., and de Streel, A. (2020). Making data portability more effective for the digital economy: Economic implications and regulatory challenges of the portability of personal data in the digital economy. CERRE Policy Report.

access to their data. Consumers may have to wait up to a month or longer to receive the portable data from their current provider and may face constraints regarding the frequency of these requests. Moreover, often consumers do not want to immediately switch to a new provider completely, but multi-home between providers first.[342] In this case consumers may not switch if they have to terminate their contract with the other provider in order to exercise their right to data retrieval (as under Article 16 DCD), and also here a much more frequent porting of data than what is provided by Article 20 GDPR would be desirable.

Thirdly, given the novelty of the right to data portability, firms also raise **legal concerns and uncertainties** that might arise when including data in data portability requests and when accepting data from other providers. This includes potential conflicts of rights, especially regarding the porting of data provided by the data subject on other data subjects (e.g., address books, or pictures in which other people are tagged). But legal concerns also arise with respect to liability issues, such as who is responsible if data is lost or modified in the transfer process. The White Paper on Data Portability by Facebook (2019)[343] summarizes these legal concerns well. However, some of those concerns can be addressed with the current legal rules. In any case, in order to encourage more to be included under the scope of data portability and firms to be more willing to import data, especially in the context of the digital economy, **more legal certainty and guidance** would be welcomed. Moreover, there may be a role for a regulatory testbed, where innovative start-ups accepting ported data, could work more closely together with the privacy regulator in order to develop legally sound and economically viable solutions. However, as this report is focused on economic and technical aspects of data sharing, but not with the legal and governance issues, we refer to Krämer et al. (2020) as well as the companion report by Feasey and de Streel (2020) for policy measures on this issue.

Fourthly, as the sharing of personal data, facilitated by existing and future regulations, is envisaged to become an important pillar of consumer empowerment as well as competition and innovation in the digital economy, **more attention should be paid to the role of PIMS**. They could provide centralised management of users' privacy settings and consented data flows; ideally aggregating relevant information across the various digital services that a consumer is using, and being able to change settings across several services as needed. In this sense, PIMS would provide a dashboard of dashboards for users' privacy settings. However, the development of PIMS is still in its infancy.[344]

Moreover, there are also economic concerns over whether privately financed 'neutral' PIMS, which act purely on the behalf of consumers, could ever find a sustainable business model (see Section 5 in Krämer et al., 2020, for a detailed discussion).

Several PIMS that set out to monetize personal data on behalf of their users has failed in the recent past.[345] Paying users for their data also gives rise to an ethical issue, as such practice would quickly reveal that the data of some users is more valuable than the data of others. Moreover, the social externality of data (Acemoglu et al. 2019[346]; Bergemann, et al. 2020[347], see also Section 5.2.1), also means that a data intermediary can acquire information about users at relatively little costs. This fundamentally undermines the idea that 'data ownership' of one sort or another empowers

[342] In this sense also Crémer et al., 2019, p.82.
[343] Facebook (2019). Charting a Way Forward: Data Portability and Privacy (September 2019). Available at: https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf
[344] See Section 3 in Krämer, Senellart and de Streel (2020).
[345] An example is Datacoup (https://www.datacoup.com), which, according to Wikipedia, offered each user up to USD 5 per month, and in the beta phase up to USD 8 per month in return for access to user accounts of various social networks such as Facebook and LinkedIn, as well as to debit and credit card transactions. However, in November 2019 Datacoup announced its users that it is closing down, and had actually never sold any of their data up to this point. Instead, all payments had been made from the Datacoup treasury account. Other examples of PDBs are people.io (http://people.io), whch seems to face similar issues as Datacoup, Datum (https://www.datum.org), where data can be sold in return for cryptocurrency only, ItsMyData (https://itsmydata.de/?lang=en), which plans to pay consumers in the future, but does not do so yet), and Wibson (https://wibson.org), where users can earn tokens that shall be redeemable in a yet to be launched marketplace. Even the large telecom operator Telefonica has announced a personal data space with the possibility to be redeemed for data in 2017 (see https://www.ft.com/content/3278e6dc-67af-11e7-9a66-93fb352ba1fe) as part of their AI-project 'Aura', but the project seems to focus on the Ai functionalities now. Note that 'Aura' now refers to another, unrelated project at Telefonica, a personal digital assistant.
[346] Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). Too much data: Prices and inefficiencies in data markets (No. w26296). National Bureau of Economic Research. Available at: https://economics.harvard.edu/files/economics/files/acemoglu_spring_2020.pdf
[347] Bergemann, D., Bonatti, A., & Gan, T. (2020). The economics of social data. Cowles Foundation Discussion Paper No. 2203R. Available at: https://ssrn.com/abstract=3548336

consumers to receive a 'fair' and significant remuneration for their data, and hence whether users would ever transfer personal data to a PIMS for financial gain.

However, PIMS and other technical solutions for standardised data exchange are still in their infancy.[348] Noteworthy open-source non-profit **projects are Solid and the Data Transfer Project (DTP)**. The Data Transfer Project is a technology initiative that was launched in 2018 and is supported, among others, by Apple, Facebook, Google, Microsoft and Twitter. The main outcome of this initiative is the development of a specification of an open-source platform for data transfer. Though these five companies are nominally involved, the project inherits from Google's former *Data Liberation Front*, and Google is by far the main contributor to the DTP platform. Both Solid and DTP are, when compared to other successful open source projects, still at a very early stage of development and have progressed relatively little in the recent past.

Finally, there is **limited evidence that data portability is widely used to date**. The root of this seems to be a classic chicken-and-egg problem. Not at least for the reasons given above, currently very few providers to indeed accept ported data from users. If data is imported, it is often not done via the data set that a user has exported following a data portability request, but rather through existing APIs or other workarounds. In reverse, this means there is a lack of use cases for consumers to exercise their right to data portability. We believe that **more continuous and standardised data portability is key to overcoming this chicken-and-egg problem**. Moreover, the experience from telecom markets (number portability) shows that portability became widely adopted when the consumer merely needs to give consent, but the (technical) details of the exchange are deliberated by the sending and receiving data controllers directly according to some standardised process. The experience from other industries, foremost the Open Banking Order in the UK, highlights that third-parties often do see a value in importing data, and that data importing becomes more likely when standards are in place that allows for continuous imports of data. In the case of Open Banking, after a slow start, there has been a continuous increase in both the number of third-parties accessing the available APIs as well as in the number of API calls being made.[349]

Taken together, Krämer et al (2020) see scope for **improvement in the context of personal data portability in three areas: (I) effective enforcement of the current legal framework, (ii) a new right for continuous data portability, tailored for the digital economy, and (iii) enabling PIMS through standards**. Here we focus merely on the role of continuous data portability as a potential data-sharing remedy.

5.3.2 Continuous data portability

To empower users to switch and multi-home digital service, and to facilitate real-time and continuous access of third-parties to keep user data, we argue that it is **necessary to introduce new legislation** which enables consumers to transfer their data (as under Article 20 GDPR) and their non-personal data (as under Article 16 DCD) in a timely and frequent manner from their existing digital service provider to another provider, at any given point in time. This is what we refer to as 'continuous data portability'.

This is not an entirely new policy proposal. It is in a similar spirit as the "Smart Data" initiative in the UK, which, however, is initially limited to regulated industries, beginning with the Open Banking, but also seeks to include digital markets in the future.[350] Similar steps are being taken under the new Consumer Data Right (CDR) in Australia. The policy proposal also relates to the recently adopted European data strategy, who recognizes that the "absence of technical tools and standards" makes the exercise of data portability burdensome.[351] Indeed, even several of the largest tech firms recently expressed their efforts to give users more control over their data and privacy.[352] Facebook CEO Mark Zuckerberg explicitly urged governments for more regulation and identified data portability as one

[348] Section 4 in Krämer, Senellart and de Streel (2020) discusses in detail why PIMS struggle to find a sustainable business model, which is also a reason for their slow development.
[349] See https://www.openbanking.org.uk/providers/account-providers/api-performance/
[350] See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/808272/Smart-Data-Consultation.pdf
[351] Communication from the Commission of 19 February 2020, A European strategy for data, COM(2020) 66, p.10.

[352] See, e.g., https://eandt.theiet.org/content/articles/2020/01/google-ceo-backs-gdpr-says-privacy-should-not-be-a-luxury/

of four areas where such action should be taken.[353] The envisioned regulation on continuous data portability would be a step in this direction.

Based on Krämer et al (2020), we propose the following principles for continuous data sharing:

Principles for the scope and implementation of continuous data portability to facilitate consent-based sharing of deep user data with third-parties

1. Only **raw user data** (observed and volunteered) may be subject to continuous data sharing, but not derived insights from such data
2. Consumers must be able to give their **consent on a fine-granular level** regarding which data is to be transferred. All-or-nothing transfers are often not necessary and would create more transaction costs, both technically (e.g., network load, space requirements) as well as economically (larger privacy concerns). They would also run counter the legal requirements of data minimisation under GDPR; firms shall not influence consent or discontent by offering commercial incentives or disincentives.
3. Data should be able to be **shared directly between firms** when consumers have consented to this. Continuous data portability should be possible without any additional infrastructure at the consumer end. However, this does not preclude the possibility that users employ PIMS to store data or to facilitate this process.
4. Relatedly, the nature and scope of the data ported should be very **clearly communicated** to consumers, in plain language
5. The data transfer needs to be **secure**, minimizing risks for data leakage to parties not involved in the transfer, data modification or loss of data
6. Where possible **open standards and protocols** should be used, which are free to use and transparent for developers
7. **APIs need to be available with high reliability and performance**. They should have the same performance and reliability as the interfaces that consumers otherwise use to interact with the digital service provider (as in the PSD2).

Some additional comments are in order. The principles suggest that the **scope** of data covered under a continuous data portability regulation should be exactly as under Art. 20 GDPR and Art. 6 DCD. The policy proposal is therefore not to widen the scope of data access rights that users are already entitled to, but to increase the effectiveness and immediacy in which they can exercise these rights. Whereas GDPR is a horizontal regulation that applies to all firms, not just in the context of the digital economy, continuous data portability is specifically targeted at the technical possibilities and economic realities in the digital economy.

On a similar note, the provisions in the **GDPR on purpose limitation, data minimisation and data portability create particularly strong tensions** in the context of the digital economy, where data is always processed by automated means and every click is potentially recorded. Specifically, there is an ongoing legal discussion to what extent observed data (as opposed to volunteered data) should be included in the right to data portability. In its interpretative guidelines, the European Data Protection Board (EDPB) takes a broad view and suggests that both observed and volunteered data, however not inferred data, should be included in the scope of data portability.[354] If this interpretation is followed, detailed behavioural data (e.g., clickstreams, viewing and purchase histories, etc) should be subject to data portability on the request of the user. This interpretation would also be supported and encouraged by our analysis here.
Moreover, we echo the proposals made in the Furman Report (2019, pp. 71-74) and elsewhere that ***open standards*** **and protocols** should be used. The development of standards and technical solutions can be built on existing open-source projects such as the Data Transfer Project or Solid. Of course, the devil is in the detail and implementing this involves challenges, as the implementation of PSD2/Open Banking or cloud-based services like IaaS and SaaS have shown. Given the demonstration project of DTP and Solid, there does not seem to be a compelling technical reason

[353] See https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html
[354] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, WP242 rev.01, p. 10

why this is not feasible in a wider context. It is also to be expected that, once standards are defined and APIs are available, there will be a significant effort from the open-source community to provide import and export adapters between various services. Although this process should be industry-led, there should be a time deadline after which the progress and implementation status is evaluated by the Commission. If no sufficient progress has been made by means in establishing standards and operational interfaces within a specified period, there may be a need for stronger governmental intervention or guidelines to ensure progress is made and neutrality of interests are warranted. For example, in Open Banking the major banks were required to constitute an independent trustee to develop standards. In the case of PSD2, relatively detailed technical provisions were adopted by the Commission based on the participatory work done at the European Banking Authority. Similar case-by-case provisions are also done in the Australian Consumer Data Right (CDR) initiative.[355] The ultimate option of last resort is to enact a public standards organization to achieve this end. For example, the Australian government has given a legal mandate the Data Standards Body to develop standards for data access and portability.[356] It works in close collaboration with the competition authority and the data protection authority.

[355] See https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0
[356] See https://consumerdatastandards.org.au

06

CONCLUSIONS

6 Conclusions

6.1 Summary and main results

In this report, we have provided an in-depth analysis of the role of data for competition and innovation in digital markets and discussed various data-related remedies that could be used to enhance competition and innovation in data-driven digital markets.

6.1.1 The economic value of data

To assess the role of data in today's digital markets, the report first investigated **the role of data for service quality and competition in three key digital markets**: general search, e-commerce and media. With respect to ***general search***, we described the **economic value that is derived from the collection of web index data, search query data, data on user behaviour and individual user data**. While search index data is the core input for web search engines, this data is publicly available and does not present a data bottleneck per se. Large scale search query data and behavioural use data, such as the users' interaction with the search results, are equally important for the quality of a search engine because they allow to significantly improve the matching and ranking of search results, especially for new and rare queries. However, such user data is not publicly available and proprietary to the search engine provider. Moreover, individual user data, which is often collected also outside of search in relation to other services, is used to tailor search results to individual-level contexts and preferences. Besides, we evaluated the data requirements for local search and characterised how search engines can leverage their position as an information intermediary to incentivise third-party businesses to provide and create additional proprietary data. In local search, individual-level context data is even more important to infer a user's intent. We also examined search advertising and illustrated how observed data on search queries and user data can be monetized by improving the effectiveness of search advertising.

Concerning ***e-commerce***, we described the collection and use of data for demand forecasting, ancillary platform services for third-parties and personalised recommendations. We focused on the use of data for recommendation systems, which are especially important for online marketplaces that offer large product catalogues. **Aggregate behavioural user data (product interactions, purchases) is useful to improve demand forecasting**, which in turn is crucial for product portfolio decisions and for improving the efficiency of operations. Marketplace operators can also observe **data from transactions of third-parties on the platform**, and through ancillary platform services, also off the platform. Moreover, we reviewed the role of several data inputs, such as consumer-product data, data on user behaviour, individual user data and product data for the accuracy of personalised recommendations. Fine-granular data on user behaviour improves recommendation accuracy significantly, and product data can help to overcome the cold-start problem of recommendation systems. Data on user behaviour on other services can be used to infer more general preferences and new similarity relationships between users. However, our analysis also highlights that the **business value of data inputs can materialise along different business dimensions**. For example, improved recommendation quality can increase customer satisfaction and thus improve long-term profitability due to reduced churn of customers, but recommendations may also increase conversion rates by improving users' discovery of items with better product fits.

Concerning digital ***media platforms***, we reviewed the collection and use of data for increasing the appeal of content to users through personalisation, service improvement and algorithmic content moderation. We show that personalisation of content requires matching identifiable personal data to non-personal data about the content, whereas aggregate personal and non-personal data goes into service improvement including consumer protection measures. However, our focus is on the use of data for advertising purposes. Advertising generally is a predictions game that requires a continuous feed of aggregate user data in the planning and measuring of campaigns. Data is also especially relevant for the sale of targeted advertising in the case of advertising-financed media platforms. Segment-based targeting and all behavioural advertising require timely, accurate and deep identifiable personal data. The **greatest economic value here lies in first-party data and channels for extracting campaign data**.

As illustrated across these three case studies, data is at the core of most digital services today. For all markets surveyed, we conclude that **more data, especially more data on user behaviour, will gradually improve the quality of the digital service, albeit at a decreasing marginal rate, and allow the firms to generate higher economic benefits along various business**

value dimensions. This positive feedback loop is what characterizes data-driven markets and leads to data-driven network effects that create high entry barriers for firms that do not have access to such data. Although in all three markets it is feasible to enter with a basic service that does not use (behavioural) data, such a service would often be insufficient to attract users and to grow a viable customer base.

Concerning scale and quality advantages, the considered case studies demonstrate **that data is often created as a by-product of consumers' usage of a service**. The **scale of operations** therefore directly increases the ***breadth of data*** that is available to a firm. Breadth related to how many users are contained in the available data and, thus, how representative the data is for the total population of consumers. We show that empirical evidence points to positive but diminishing returns from broader data sets. When collected data can be associated with individual users, this increase the ***depth of data***, i.e., the average length of a user profile increases and more information per user becomes available. Empirical studies show that longer user profiles may play an important role with regard to the economic benefits from increasing data scale. On the one hand, additional user information may yield direct improvements with respect to the performance of algorithms. As in the case of broader data sets, these improvements are found to be positive, but diminishing with larger depth. On the other hand, more user data may at the same time reinforce the benefits from broader data sets. This is, because the user data does not only benefit the performance of algorithmic tasks targeted at this individual user but also improves the performance of tasks targeted at other users that are identified as similar users, based on the individual-level data. This may give rise to data-driven network effects even in the absence of increasing returns to scale.

Next to the scale of data sets, the **quality of data** significantly influences the economic value of data that can be extracted. Moreover, quality requirements will determine the competitive ramifications if firms have unequal access to data. Specifically, the timeliness of data is important to consider, as consumers' preferences change over time and new relevant items such as products or websites appear in the respective business context. In cases where data outdate quickly, the incumbency advantage of directly observing user behaviour will be especially relevant.

Finally, we highlighted that the analysis of data-driven competitive advantages must consider the **complementary inputs** that are required for the collection and processing of data. In particular, this comprises computing and storage infrastructure, skilled human resources and algorithms.

6.1.2 Data-driven theory of harm and policy objectives

We then assessed and clarified the underlying **theory of harm** for data aggregation and data exclusiveness. At its root is the presence of data-driven network effects, which likely leads to the **tipping of a market**, such that only one dominant provider prevails, and which creates high entry barriers. In a tipped market, innovation incentives of both the incumbent and potential entrants are likely to be lower than in a competitive market. Moreover, data-driven network effects also give rise to a **domino-effect**, which allows data rich incumbents to enter into adjacent markets, thereby increasing their ability to collect data even more. This is facilitated by envelopment strategies, whereby existing services are bundled with the new service. **Particularly, ancillary data services**, such as digital identity management services or financial transaction services may be viewed with scepticisms, because they allow the collection of additional data. However, in this case, providers of such ancillary data services are not competing and innovating in these markets themselves. Additional harms for data access may arise in the context of **vertical relationships**, e.g. when firms are providing both a platform and act as a provider on the platform.

Finally, there is also increasing evidence that data-driven network effects and associated entry barriers harm venture capital **for innovative start-ups** that seek to contest the business model of data-rich incumbents. The reason is that such start-ups often find themselves in a 'kill zone', where they are driven out of the market, either through the incumbent's lower marginal costs of innovation (caused by data-driven network effects) or through acquisition.
However, data-driven network effects also bear **inherent efficiencies** that must be considered before any policy intervention. Realising economies of scale and scope in data aggregation, which create entry barriers on the one hand, also generally benefit consumers on the other hand, because they allow to identify and develop products and services that cater to a consumer's individual needs and preferences and create efficiencies that would not have been able otherwise.

We, therefore, argued that **contestability in the narrow sense**, i.e., replacing the incumbent by a more efficient entrant in a process of 'creative destruction', is **neither a realistic nor necessarily a desirable** policy objective. Even if access to (user) data is facilitated through policy interventions, a significant data advantages will remain with the incumbent, not the least because deep personal data is not sharable without a user's consent. Hence, we suggest that policy makers should focus on **enabling *niche entry and niche growth*** and a ***level playing field* for competitors in new and emerging markets**.

In this context, we suggested that the **discussion of *essential data* may be futile** because 'essential data' in the meaning of the essential facilities doctrine often does not exist. Market entry is possible without access to proprietary behavioural user data and can be based purely on publicly or otherwise commercially available data. However, in practice ***access to such behavioural data* would be necessary for many instances to offer a competitive service or to develop data-based innovations in other domains.**

6.1.3 Possible data access remedies

At first, we reviewed different **data remedies that aimed at limiting the collection of user data** with respect to their technical feasibility and the economic trade-offs involved. These remedies included data siloing (i.e., preventing aggregation of data originating from different services), shorter data retention periods, prohibiting incumbents from buying into default settings, line of business restrictions, and privacy-enhancing technologies. The general problem with these sets of remedies is that they seek to **achieve a more level playing field in the digital economy by breaking the data-driven network effects** of the incumbents. This is associated with diminishing the efficiency of the incumbent and also the ability to create value from data more generally. Although data minimisation is a value in its own right from a privacy perspective, our assessment here was mainly based on economic rationales in view of facilitating market contestability and niche entry. From a mere economic perspective, we argued that **many of these remedies would not be effective in fostering competition and entry in digital markets**. However, line of business restrictions, including vertical separation may be considered by policy makers under very specific conditions, and as a remedy of last resort if data sharing remedies should prove to be ineffective. In particular, we suggest that policy makers should **consider the possibility to restrict the use of ancillary data services** by incumbents, in so far as they allow to track user behaviour across the entire Internet, e.g. identity management services, financial services or web analytics services. Such services make it very difficult for consumers to truly control which firm they are providing their user behaviour data, and they undermine exclusive data advantages of niche competitors, which may help them to grow and scale. Moreover, such ancillary data services may often be similarly provided by independent third parties, and with relatively little, if any, efficiency losses. Finally, **privacy-enhancing technologies should generally also be part of the regulatory toolkit**, but must be tailored to the specific use case and must generally be accompanied with other remedies.

Next, we discussed the application and scope of **data sharing remedies that aim at providing access to broad user data**. We argue that, to preserve innovation incentives, only **raw user data** (observed and volunteered) may have to be shared. Moreover, only data that was created as a **by-product** of consumers' usage of a dominant service should be within the scope of mandated data sharing (e.g., search queries or location data); but not (volunteered) user data that represents the essence of the service itself (e.g., posts on a social media site). The line may be sometimes difficult to draw in practice, but it is important to make this distinction because otherwise legitimate existing business models may be destroyed and innovation incentives will be unduly harmed. Data should also generally be made available through **standardised interfaces (APIs) in real-time** and continuously.

The most difficult part will be to **balance privacy concerns** with maintaining enough level of granularity in the data, such that it is valuable for data-driven innovations by third-parties. We survey several technical and institutional means that can facilitate this balancing act and prevent de-anonymisation of shared data sets. Within limitations, we entertain the idea that a data trust and data sandboxing (at a data trust) may be feasible if confined to subsets of the data to be shared, particularly with a focus on recency, and if confined to a few select algorithms that may be trained at any given time. The EuroHPC, a European collective effort to create a supercomputing ecosystem, may be the technical host to such a data trust. Furthermore, we also see some merit in the proposal to declare deliberate de-anonymisation efforts illegal under European law.

We also made specific proposals to advance the debate on **broad user data sharing in the context of our three case studies**. Concerning ***search***, we suggest three categories of data from which data access requests should be considered: i) data on the search query, ii) data on the search results page, and iii) data on the user. Generally, complex trade-offs are to be considered and we suggest that mandated access to data needs to be done on a case-by-case basis and requires a vetting procedure of the data access seeker by the regulatory authority. This will likely come alongside with additional responsibilities and safeguards for the data recipient. At the same time, a less detailed, highly anonymised data set should be made publicly available without prior vetting.

Concerning ***e-commerce***, we are sceptical that any mandated sharing of broad user data would be warranted, albeit the transparency of data use as well as the detail and mobility of information that is already provided by platforms could be improved. In particular, competition in and for e-commerce markets is already intense, and not only focused on data use but also on price. Also in view of the increased e-commerce related activities of Google and Facebook, regulatory forbearance with respect to mandated data sharing seems to be in order for the time being.

In the context of ***advertisement-supported social media platforms***, the most contentious issues related to the access to ad campaign data and user interaction with advertisements. Here data access restrictions are often of contractual rather than technical nature, and also subject to ongoing investigations, e.g. by the CMA in the UK. While there may be legitimate interests on both sides to require and deny access to more user data, we suggest that users may be allowed to opt-in into the sharing of their behavioural data with content creators and/or advertisers. This may alleviate privacy concerns on the one hand and raise and allow content creators to pledge their loyal audience to trust them with their data.

Finally, we discussed how **access to 'deep' raw user data** can be facilitated by strengthening consumer rights above and beyond their existing data portability right under Article 20 GDPR. In particular, we suggest that in several cases competition and innovation would benefit if firms were obliged to provide consumers with the possibility to consent to continuous, real-time data portability. The scope of data to be transferred should be identical as under GDPR Art. 20. However, to date more legal certainty is needed with respect to the precise scope of Art. 20 GDPR with respect to observed (user behaviour) data. Generally, as in the case of mandated sharing of broad user data, only raw user data (volunteered and observed) should be subject to data portability. Besides, consumers need to consent to every such continuous transfer. Continuous data portability should be made possible through standardised APIs, enabling both business-to-business data transfers, but also the use of Personal Information Management Systems (PIMS). Demonstration projects like the Data Transfer Project and Solid exemplify that such continuous data portability is feasible from a technical perspective. However, mandating continuous data portability will require policy makers also to facilitate the setting of and agreeing on (open and secure) standards for data transfers, and consumer consent.

6.2 Which markets and firms should be subjected to a data-sharing regulation?

As of now, we have left the question open which firms and markets should be subjected to the data sharing remedies that were proposed. This important question is the scope of the companion CERRE study by Feasey and de Streel (2020), which explores a corresponding governance structure for data sharing in digital markets.

Generally, it is argued that **data sharing should be part of a regulatory toolkit** that could be required under specific conditions considering the specifics of the case. As we have highlighted here, there is generally **no one-size-fits-all approach** for data sharing and other data-related remedies. Moreover, the different remedies that we suggested, ranging from a line of business restrictions for ancillary services to mandated broad data sharing and continuous data portability, also differ in their impact on the regulated firm over and beyond the status quo for unregulated firms. Feasey and de Streel (2020) therefore suggest that the **thresholds for intervention should be set accordingly**.

They depart from the notion that some digital incumbents will be designated with a **'significant market status' (SMS)** by a competent authority. This status is derived on a case by case basis and does not just relate to a firm's dominant role in collecting user data (on which we have focused here), but also considers market power and 'intermediation power' in digital markets more generally. SMS likely comes with heightened obligations for fair and non-discriminatory conduct over and beyond

those prescribed by the P2B regulation. However, SMS may not necessarily include data sharing obligations per se, at least not with respect to bulk sharing of broad user data.

From this set of firms with SMS, a ***smaller* subset of firms would be mandated to share broad user data**, as was suggested in Section 5.2, based more specifically on the role of that firm in collecting relevant user data. This set (or a possibly even smaller set) of firms may also be subject to limitations in data collection, specifically the line of business restrictions with respect to ancillary data services, as suggested in Section 5.1.

In reverse, a ***larger* set of firms than those with SMS could be subjected to implement continuous data portability**. The reasoning here is that data portability per se is not a new right for consumers but immediately builds on existing rights of data portability under GDPR and DCD. Moreover, continuous data portability is meant not only to facilitate competition and innovation by third-parties but also to empower consumers in the digital economy more generally. The main trade-off, therefore, occurs with respect to the proportionality of an obligation to implement continuous data portability, particularly with respect to smaller and emerging firms. This is especially so because the expressed policy objective is to facilitate niche entry in digital markets and to allow less data-rich firms to grow and scale.

In addition to a governance structure that identifies which firms are obliged to sharing more data, it will also be necessary to develop a **governance structure for the supervision of the sharing process** itself. This requires a governance body that is responsible for the vetting of firms requesting access to certain data sets, and for determining the scope of data access (categories and detail of data to be shared) in these cases. In this context, the governing body will also be responsible for determining conditions for fair and non-discriminatory terms of access to the shared data sets. This may include the determination of an access fee to be paid to compensate the access provider for the transaction costs involved in continuously providing data. But it may also involve a decision whether the data is shared by the access provider directly, or through a data trust, as detailed in Section 5.2.3. In the latter case, the governing authority may also oversee the operations of the data trust. Finally, the governance body should also oversee the timely progression of the standards-setting process that is needed to offer coherent technical interfaces for continuous data sharing across different data providers.

REFERENCES

References

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). *Too much data: Prices and inefficiencies in data markets* (Discussion Paper DP14225). Centre for Economic Policy Research. https://repec.cepr.org/repec/cpr/ceprdp/DP14225.pdf

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, *24*(4), 956-975. https://doi.org/10.1287/isre.2013.0497

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, *29*(1), 84-102. https://doi.org/10.1287/isre.2017.0703

Adshead, S., Forsyth, G., Wood, S., & Wilkenson, S. (2019). *Online Advertising in the UK.* UK Department of Media Culture and Sport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/777996/Plum_DCMS_Online_Advertising_in_the_UK.pdf

Aggarwal, C. C. (2016). *Recommender Systems*. Springer International Publishing.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. *The Quarterly Journal of Economics*, *120*(2), 701-728. https://doi.org/10.1093/qje/120.2.701

Agichtein, E., Brill, E., & Dumais, S. (2006, August). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 19-26). ACM. https://doi.org/10.1145/1148170.1148177 .

Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, *14*(2), 37-48. https://doi.org/10.1145/2481244.2481250

Amatriain, X., & Basilico, J. (2015). Recommender systems in industry: A Netflix case study. In *Recommender Systems Handbook* (pp. 385-419). Springer.

Anderson, S. P., & Jullien, B. (2015). The advertising-financed business model in two-sided media markets. In *Handbook of Media Economics* (Vol. 1, pp. 41-90). North-Holland.

Ann. (2016, January 26). *How Amazon uses its own cloud to process vast, multidimensional datasets*. DZone. https://dzone.com/articles/big-data-analytics-delivering-business-value-at-am

Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, *40*(2), 131-145. https://doi.org/10.1509%2Fjmkr.40.2.131.19224

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendation Systems. *Journal of Marketing Research*, *37*(3) 363-375. https://doi.org/10.1509%2Fjmkr.37.3.363.18779

Argenton, C., & Prüfer, J. (2012). Search engine competition with network externalities. *Journal of Competition Law and Economics*, *8*(1), 73-105. https://doi.org/10.1093/joclec/nhr018

Arrow, K. J. (1962). Economic Welfare and the Allocation of Resources for Invention. *The Rate and Direction of Inventive Activity*, 609–626. https://doi.org/10.1515/9781400879762-024

Australian Competition & Consumer Commission (ACCC). (2020). *Consumer data right (CDR)*. https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2018). *The impact of big data on firm performance: An empirical investigation* (NBER Working Paper No. 24334). National Bureau of Economic Research. https://www.nber.org/papers/w24334

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019, May). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings* (Vol. 109, pp. 33-37). https://doi.org/10.1257/pandp.20191000

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing* (pp. 462-474). Springer.

Barbaro, M. and Zeller, T. (2006, August 9). *A face is exposed for AOL searcher No. 4417749*. New York Times. https://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0

Barker, A. (2020, February 26). *'Cookie apocalypse' forces profound changes in online advertising*. Financial Times. https://www.ft.com/content/169079b2-3ba1-11ea-b84f-a62c46f39bc2?shareType=nongift

Barwise, P., & Watkins, L. (2018). The evolution of digital dominance: How and why we got to GAFA. In M. Moore, & D. Tambini (Eds.), *Digital dominance: The power of Google, Amazon, Facebook, and Apple* (pp. 21-49). Oxford University Press.

Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, *52*(1), 1-37. https://doi.org/10.1007/s10462-018-9654-y

Baumol, W., Panzar, J., & Willig, R. (1982). Contestable markets and the theory of industry structure. Harcourt Brace Jovanovich.

Bergemann, D., Bonatti, A., & Gan, T. (2020). *The economics of social data* (Cowles Foundation Discussion Paper No. 2203R). https://arxiv.org/pdf/2004.03107.pdf

Bershidsky, L. (2019, March 13). *Breaking up big tech is too scary for Europe*. Bloomberg. https://www.bloomberg.com/opinion/articles/2019-03-13/breaking-up-amazon-facebook-and-google-is-too-scary-for-europe

Bleier, A., & Eisenbeiss, M. (2015). The importance of trust for personalized online advertising. *Journal of Retailing*, *91*(3), 390-409. https://doi.org/10.1016/j.jretai.2015.04.001

Blumenthal, M. (2018, October 22). Reserve with Google makes its way into the 3-pack SERP on Google. *Understanding Google My Business & Local Search.* http://blumenthals.com/blog/2018/10/22/reserve-with-google-makes-its-way-into-the-3-pack-serp-on-google/

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, *46*(3), 363-376. https://doi.org/10.1080/00913367.2017.1339368

Bourreau, M., De Streel, A., & Graef, I. (2017). *Big Data and Competition Policy: Market power, personalised pricing and advertising.* Centre on Regulation in Europe (CERRE). https://www.cerre.eu/sites/cerre/files/170216_CERRE_CompData_FinalReport.pdf

Bourreau, M., & De Streel, A. (2019). Digital conglomerates and EU competition policy. *Available at SSRN 3350512*. https://dx.doi.org/10.2139/ssrn.3350512

Bourreau, M., & De Streel, A. (2020). *Competition & innovation effects and EU merger control* (CERRE Issue Paper). Centre on Regulation in Europe (CERRE). https://www.cerre.eu/sites/cerre/files/cerre_big_tech_acquisitions_2020.pdf

Broughton Micova S., Jacques, S. (2014). *The playing field for audiovisual advertising: What does it look like and who is playing*. Centre on Regulation in Europe (CERRE). https://www.cerre.eu/sites/cerre/files/cerre_playingfieldaudiovisualadvertising_2019april_0.pdf
Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. *Sloan Management Review*, *47*(4), 67-71.

Brynjolfsson, E., Hu, Y., & Smith, M. D. (2010). Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research*, *21*(4), 736-747. https://doi.org/10.1287/isre.1100.0325

Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, *57*(8), 1373-1386. https://doi.org/10.1287/mnsc.1110.1371

Brynjolfsson, E., & McElheran, K. (2016). *Data in action: data-driven decision making in US manufacturing* (Working Paper No. 16-06). US Census Bureau Center for Economic Studies. https://dx.doi.org/10.2139/ssrn.2722502

Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, *106*(5), 133-39. https://doi.org/10.1257/aer.p20161016

Bundesgerichtshof (2020, June 23). *Bundesgerichtshof bestätigt vorläufig den Vorwurf der missbräuchlichen Ausnutzung einer marktbeherrschenden Stellung durch Facebook* [Press Release]. https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2020/2020080.html;jsessionid=F02FBF1A27F70DFD5DD8DC318EFB6C59.1_cid368?nn=10690868

Bundeskartellamt. (2019, February 2). *Bundeskartellamt prohibits Facebook from combining user data from different sources* [Press Release]. https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html

BusinessWire. (2016, June 14). *How many products does Amazon actually carry? And in what categories?* https://www.businesswire.com/news/home/20160614006063/en/Products-Amazon-Carry-Categories

Cantador, I., Fernández-Tobías, I., Berkovsky, S., & Cremonesi, P. (2015). Cross-domain recommender systems. In *Recommender Systems Handbook* (pp. 919-959). Springer. https://doi.org/10.1007/978-1-4899-7637-6_27

Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J. T., Chen, E., & Yang, Q. (2009, July). Context-aware query classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-10). ACM. https://doi.org/10.1145/1571941.1571945

Carmi, E., Oestreicher-Singer, G., Stettner, U., & Sundararajan, A. (2017). Is Oprah Contagious? The Depth of Diffusion of Demand Shocks in a Product Network. *Management Information Systems Quarterly*, *41*(1), 207-221.

Chapelle, O., & Chang, Y. (2011, January). Yahoo! learning to rank challenge overview. *Proceedings of the Learning to Rank Challenge*, (pp. 1-24).

Chen, Y., & Canny, J. F. (2011). Recommending ephemeral items at web scale. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 1013-1022). https://doi.org/10.1145/2009916.2010051

Chen, J., & Stallaert, J. (2014). An Economic Analysis of Online Advertising Using Behavioral Targeting. *MIS Quarterly, 38*(2), 429-A7.

Chiou, L., & Tucker, C. (2017). *Search engines and data retention: Implications for privacy and antitrust* (NBER Working Paper No. 23815). National Bureau of Economic Research. https://www.nber.org/papers/w23815

Choi, J. P., & Stefanadis, C. (2001). Tying, investment, and the dynamic leverage theory. *RAND Journal of Economics*, 52-71. https://doi.org/10.2307/2696397

Chong, A. Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, *55*(17), 5142-5156. https://doi.org/10.1080/00207543.2015.1066519
Clark, J. (2015, October 26). *Google turning its lucrative web search over to AI machines.* Bloomberg. https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines

Claussen, J., Peukert, C., & Sen, A. (2019). *The Editor vs. the Algorithm: Targeting, Data and Externalities in Online News* (Working Paper). https://dx.doi.org/10.2139/ssrn.3399947

Clement, J. (2020, April 28). *Mobile share of U.S. organic search engine visits 2013-2020*. Statista. https://www.statista.com/statistics/297137/mobile-share-of-us-organic-search-engine-visits/

Commission Nationale de l'Informatique et des Libertés (CNIL). (2019, January 21). *The CNIL's restricted committee imposes a financial penalty of 50 million euros against Google LLC*. https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc

Competition & Markets Authority (CMA). (2019a). *Online platforms and digital advertising market study: Observations on the CMA's interim report*. https://assets.publishing.service.gov.uk/media/5dfa0580ed915d0933009761/Interim_report.pdf

Competition & Markets Authority (CMA). (2019b). *Online platforms and digital advertising. Market study interim report.* https://assets.publishing.service.gov.uk/media/5dfa0580ed915d0933009761/Interim_report.pdf

Competition & Markets Authority (CMA). (2019c). *Online platforms and digital advertising. Market study interim report – Appendix L.* https://assets.publishing.service.gov.uk/media/5df9efa2ed915d093f742872/Appendix_L_Potential_approaches_to_improving_personal_data_mobility_FINAL.pdf

Competition & Markets Authority (CMA). (2019, December 18). *CMA lifts the lid on digital giants* [Press Release]. https://www.gov.uk/government/news/cma-lifts-the-lid-on-digital-giants

Condorelli, D., & Padilla, J. (2020). Harnessing Platform Envelopment in the Digital World. *Journal of Competition Law & Economics*, *16*(2), 143-187. https://doi.org/10.1093/joclec/nhaa006

Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). *Competition policy for the digital era*. Report for the European Commission. https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 39-46). https://doi.org/10.1145/1864708.1864721

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search engines: Information retrieval in practice*. Reading: Addison-Wesley.

Davenport, T. H., Barth, P., & Bean, R. (2012). How Big Data Is Different. *MIT Sloan Management Review*, *54*(1), 22-24.

Day, M. (2019, December 31). *You're home alone with Alexa. Are your secrets safe?*. Bloomberg. https://www.bloomberg.com/news/articles/2019-12-31/you-re-home-alone-with-alexa-are-your-secrets-safe-quicktake

De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, *56*(11), 1930-1945. https://doi.org/10.1287/mnsc.1100.1233
De Cornière, A., & Taylor, G. (2019). A model of biased intermediation. *The RAND Journal of Economics*, *50*(4), 854-882. https://doi.org/10.1111/1756-2171.12298

De la Mano, M., & Padilla, J. (2018). Big Tech Banking. *Journal of Competition Law & Economics*, *14*(4), 494-526. https://doi.org/10.1093/joclec/nhz003

Department for Business, Energy & Industrial Strategy and Department for Digital, Culture, Media & Sport. (2019). *Smart data: putting consumers in control of their data and enabling innovation*. UK government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/808272/Smart-Data-Consultation.pdf

De Streel, A., & Feasey, R. (2020). *Data Sharing for Digital Markets Contestability: Towards a Governance Framework.* CERRE Policy Report. Centre on Regulation in Europe (CERRE).

De Streel, A., & Husovec, M. (2020). *The e-commerce Directive as the cornerstone of the internal market.* Department for Economic, Scientific and Quality of Life Policies at the request of the committee on Internal Market and Consumer Protection (IMCO). https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648797/IPOL_STU(2020)648797_EN.pdf

Dewey, C. (2016, May 11). *You probably haven't even noticed Google's sketchy quest to control the world's knowledge*. The Washington Post. https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/

Dezyre.com. (2018, September 18). *How Big Data analysis helped increase Waltmarts sales turnover?.* https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

Diemert, E., Meynet, J., Galland, P., & Lefortier, D. (2017). Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the ADKDD'17*, (pp. 1-6). ACM. https://doi.org/10.1145/3124749.3124752

Doyle, G. (2018). Television and the development of the data economy: Data analysis, power and the public interest. *International Journal of Digital Television*, *9*(1), 53-68. https://doi.org/10.1386/jdtv.9.1.53_1

Dziadul, C. (2019, June 5). *RTL and ProSiebenSat.1 ink addressable TV joint venture*. Broadband TV News. https://www.broadbandtvnews.com/2019/06/05/prosiebensat-1-rtl-ink-addressable-tv-joint-venture/

EBX. (2017, September 11). European media corporations agree on joint venture. *European Broadcasting Exchange*. http://ebx.tv/?page_id=269

The Economist (2017, May 6). *Fuel of the future. Data is giving rise to a new economy.* https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

The Economist. (2018, June 2). *Into the danger zone American tech giants are making life tough for startups.* https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups

Edelman, B., & Lai, Z. (2016). Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research*, *53*(6), 881-900. https://doi.org/10.1509%2Fjmr.14.0528

Egan, E. (2019). *Charting a way forward: Data portability and privacy* [White Paper]. Facebook. https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf

Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal*, *32*(12), 1270-1285. https://doi.org/10.1002/smj.935

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends in Human–Computer Interaction*, *4*(2), 81-173.

Elkahky, A. M., Song, Y., & He, X. (2015). A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 278-288). https://doi.org/10.1145/2736277.2741667

Ellis, M. (2018, October 23). The ultimate cheat sheet for taking full control of your Google Knowledge Panels. *MOZ Blog.* https://moz.com/blog/ultimate-cheat-sheet-google-knowledge-panels

Ellis, M. (2020, January 6). 2020 Local SEO success: How to feed, fight, and flip Google. *MOZ Blog*. https://moz.com/blog/2020-local-seo-success

Elmeleegy, H., Li, Y., Qi, Y., Wilmot, P., Wi, M., Kolay, S., Dasdam, A., & Chen, S. (2013). Overview of turn data management platform for digital advertising. *Proceedings of the VLDB Endowment*, *6*(11), 1138-1149. https://doi.org/10.14778/2536222.2536238

Engineering and Technology. (2020, January 22). *Google CEO backs GDPR, says privacy should not be a 'luxury'*. https://eandt.theiet.org/content/articles/2020/01/google-ceo-backs-gdpr-says-privacy-should-not-be-a-luxury/

Englehardt, S., & Narayanan, A. (2016, October). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 1388-1401). https://doi.org/10.1145/2976749.2978313

European Commission. (2016, April). *Flash Eurobarometer 439 – The use of online marketplaces and search engines by SMEs*. https://ec.europa.eu/information_society/newsroom/image/document/2016-24/fl_439_en_16137.pdf

European Commission. (2017a, May 18). *Mergers: Commission fines Facebook €110 million for providing misleading information about WhatsApp takeover* [Press Release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1369

European Commission. (2017b). *Guidelines of Article 29 Data Protection Working Party on the right to data portability* (WP 242 rev.01). https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611233

European Commission. (2018, April 26). *Study on data in platform-to-business relations*. https://ec.europa.eu/digital-single-market/en/news/study-data-platform-business-relations

European Commission. (2019a, March 20). *Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising* [Press Release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770

European Commission. (2019b, July 17). *Antitrust: Commission opens investigation into possible anti-competitive conduct of Amazon* [Press Release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4291

European Commission. (2020a). *The Digital Service Act package*. https://ec.europa.eu/digital-single-market/en/digital-services-act-package

European Commission. (2020b, February 19). A European strategy for data. *COM (2020) 66*. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

European Commission. (2020c, July 15). *Platform-to-business trading practices*. https://ec.europa.eu/digital-single-market/en/business-business-trading-practices

Evans, D. S. (2019). Attention platforms, the value of content, and public policy. *Review of Industrial Organization*, *54*(4), 775-792. https://doi.org/10.1007/s11151-019-09681-x

Facebook. (2020a, May 19). *Introducing Facebook Shops, a new online shopping experience*. Facebook Business. https://www.facebook.com/business/news/announcing-facebook-shops

Facebook. (2020b). *Advertising policies*. https://www.facebook.com/policies/ads/

Fast, V., Schnurr, D., & Wohlfarth, M. (2019). Data-driven market power: An overview of economic benefits and competitive advantages from Big Data use. *Available at SSRN 3427087*.

Feasey, R., & Krämer, J. (2019). *Implementing effective remedies for anti-competitive intermediation bias on vertically integrated platforms*. Centre on Regulation in Europe (CERRE). https://www.cerre.eu/publications/implementing-effective-remedies-anti-competitive-intermediation-bias-vertically

Federal Trade Commission (2014). *Data Brokers: A Call for Transparency and Accountability.* https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Findlay, S. and Kazmin, A. (2019, February 1). *India's ecommerce law forces Amazon and Flipkart to pull products*. Financial Times. https://www.ft.com/content/29a96ff6-2615-11e9-8ce6-5db4543da632

Fishkin, R. (2019). Less than Half of Google Searches Now Result in a Click. *SparkToro*. https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/

Fleder, D., & Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, *55*(5), 697-712.

Forbrucker Radet. (2020, January 14). *Out of control: How consumers are exploited by the online advertising industry*. https://www.forbrukerradet.no/undersokelse/no-undersokelsekategori/report-out-of-control/

Funk, S. (2006, December 11). Netflix Update: Try This at Home. https://sifter.org/~simon/journal/20061211.html

Furman, J., Coyle, D., Fletcher, A., McAules, D., & Marsden, P. (2019). *Unlocking digital competition*. Report of the digital competition expert panel for the Government of the United Kingdom. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

Geradin, D., & Katsifis, D. (2020). *Online platforms and digital market study: Observations on CMA's interim report.* https://assets.publishing.service.gov.uk/media/5e8c8a4b86650c18c6afeab5/200212_Prof._Damien_Geradin_and_Dimitrios_Katsifis_Response_to_Interim_Report.pdf

Geradin, D., Katsifis, D. and Karanikioti, T. (2020). *GDPR myopia: How a well-intended regulation ended up favoring Google in Ad Tech* (TILEC Discussion Paper No. 2020-012). https://ssrn.com/abstract=3598130

GIFT. (n.d.). Joint Tech Innovation. *Global Internet Forum to Counter Terrorism.* Retrieved from https://www.gifct.org/joint-tech-innovation/ on 2020, May 3

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 1-19. https://doi.org/10.1145/2843948
Google. (2019a). Search quality evaluator guidelines. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

Google. (2019b, July 30). *Google Maps Booking API: Overview*. https://developers.google.com/maps-booking/guides/end-to-end-integration/overview

Google. (2019c, December 9). *Google Maps Booking API: Enabling payments.* https://developers.google.com/maps-booking/guides/payments/enabling-payments

Google (2020a, March 10). *Lighthouse*. https://developers.google.com/web/tools/lighthouse

Google. (2020b). *Security Center*. https://safety.google/privacy/ads-and-data/

Google. (2020b) *Google Flu Trends*. https://www.google.org/flutrends/about/
Google Support. (2020a). *How Google sources and uses information in business listings*. Google My Business Help. https://support.google.com/business/answer/2721884?hl=en

Google Support. (2020b). *Improve your local ranking on Google.* Google My Business Help. https://support.google.com/business/answer/7091?hl=en

Google Support. (2020d). *About Local Services data*. Google Ads Help. https://support.google.com/google-ads/answer/7496727

Google Support. (2020e). *Find places you'll like*. Google Maps Help. https://support.google.com/maps/answer/7677966

Gordon, S. (2017, July 19). *Our personal data are precious – we take back control*. Financial Times. https://www.ft.com/content/3278e6dc-67af-11e7-9a66-93fb352ba1fe

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177%2F2053951719897945

Graef, I. (2015). Market definition and market power in data: The case of online platforms. *World Competition*, *38*(4), 473-505.

Graef, I. (2016). *EU competition law, data protection and online platforms: Data as essential facility*. Wolters Kluwer.

Graef, I., Wahyuningtyas, S. Y., & Valcke, P. (2015). Assessing data access issues in online platforms. *Telecommunications policy*, *39*(5), 375-387. https://doi.org/10.1016/j.telpol.2014.12.001

Green, M. (2016, June 15). What is differential privacy?. *Cryptography Engineering Blog*. https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/

Grimes, C. (2018, June 8). Our new search index: Caffeine. *Google Official Blog*. https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

Haahr, P. (2019). *Improving Search Over the Years (WMConf MTV '19)* [Video]. YouTube. https://www.youtube.com/watch?v=DeW-9fhvkLM&,

Hagiu, A., Teh, T. H., & Wright, J. (2020). *Should Amazon be allowed to sell on its own marketplace?* (Discussion Paper). https://ap4.fas.nus.edu.sg/fass/ecsjkdw/hagiu_teh_wright_may2020.pdf

Hagiu, A., & Wright, J. (2020). *Data-enabled learning, network effects and competitive advantage* (Working Paper). http://andreihagiu.com/wp-content/uploads/2020/06/Data-enabled-learning-June2020.pdf
Hajaj, N. (2015). *U.S. Patent No. 9,165,040*. Washington, DC: U.S. Patent and Trademark Office.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12. https://doi.org/10.1109/MIS.2009.36

Hao, K. (2019, November 5). *Inside Amazon's plan for Alexa to run your entire life.* MIT Technology Review. https://www.technologyreview.com/2019/11/05/65069/amazon-alexa-will-run-your-life-data-privacy/

Valdani Vicari & Associati (VVA). (2017). *Study on data in platform-to-business relations – Final report.* European Commission. https://ec.europa.eu/digital-single-market/en/news/study-data-platform-business-relations
He, D., Kannan, A., Liu, T. Y., McAfee, R. P., Qin, T., & Rao, J. M. (2017, December). Scale Effects in Web Search. In *International Conference on Web and Internet Economics* (pp. 294-310). Springer, Cham. https://doi.org/10.1007/978-3-319-71924-5_21

Heimbach, I., Gottschlich, J., & Hinz, O. (2015). The value of user's Facebook profile data for product recommendation generation. *Electronic Markets*, *25*(2), 125-138. https://doi.org/10.1007/s12525-015-0187-9

Hensel, A. (2020, January 20). Cookiepocalypse: What the death of the third-party ookie means for retailers. *Modern retail*. https://www.modernretail.co/platforms/cookiepocalypse-what-the-death-of-the-third-party-cookie-means-for-retailers/

Hinz, O., & Eckert, J. (2010). The impact of search and recommendation systems on sales in electronic commerce. *Business & Information Systems Engineering*, *2*(2), 67-77. https://doi.org/10.1007/s12599-010-0092-x

Holmstrom, B., & Roberts, J. (1998). The boundaries of the firm revisited. *Journal of Economic Perspectives*, *12*(4), 73-94. https://doi.org/10.1257/jep.12.4.73

Hong, Y., He, X., Vaidya, J., Adam, N., & Atluri, V. (2009, November). Effective anonymisation of query logs. In *Proceedings of the 18th ACM conference on Information and Knowledge Management* (pp. 1465-1468). https://doi.org/10.1145/1645953.1646146

Hölzle, U. (2012, January). *The Google gospel of speed.* Think with Google. https://www.thinkwithgoogle.com/marketing-resources/the-google-gospel-of-speed-urs-hoelzle/

Höppner, T., & Westerhoff, P. (2018, November 30). The EU's competition investigation into Amazon Marketplace. *Kluwer Competition Law Blog*. http://competitionlawblog.kluwercompetitionlaw.com/2018/11/30/the-eus-competition-investigation-into-amazon-marketplace/?doing_wp_cron=1588254455.2190229892730712890625

Hou, L., & Jiao, R. J. (2020). Data-informed inverse design by product usage information: a review, framework and outlook. *Journal of Intelligent Manufacturing*, *31*(3), 529-552. https://doi.org/10.1007/s10845-019-01463-2

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 263-272). IEEE. https://doi.org/10.1109/ICDM.2008.22

IAB Tech Lab. (2020, April). *Audience Taxonomy 1.1*. IAB Tech Lab. https://iabtechlab.com/standards/audience-taxonomy/

Internet Live Stats. (2020). *Google Band.* https://www.internetlivestats.com/one-second/#google-band

Jannach, D., & Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, *10*(4), 1-23. https://doi.org/10.1145/3370082
Jiao, J., & Zhang, Y. (2005). Product portfolio planning with customer-engineering interaction. *IIE Transactions*, *37*(9), 801-814. https://doi.org/10.1080/07408170590917011

Johnson, G. A., Lewis, R. A., & Reiley, D. H. (2017). When less is more: Data and power in advertising experiments. *Marketing Science*, *36*(1), 43-53. https://doi.org/10.1287/mksc.2016.0998

Johnson, G., Shriver, S., & Goldberg, S. (2020). Privacy & market concentration: Intended & unintended consequences of the GDPR. *Available at SSRN 3477686*. https://dx.doi.org/10.2139/ssrn.3477686

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better?. *Big Data*, *1*(4), 215-226. https://doi.org/10.1089/big.2013.0037
Katukuri, J., Könik, T., Mukherjee, R., & Kolay, S. (2014, October). Recommending similar items in large-scale online marketplaces. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 868-876). IEEE. https://doi.org/10.1109/BigData.2014.7004317

Khan, L. M. (2019). The separation of platforms and commerce. *Columbia Law Review*, *119*(4), 973-1098. https://columbialawreview.org/content/the-separation-of-platforms-and-commerce/

Kharitonov, E., & Serdyukov, P. (2012, October). Demographic context in web search re-ranking. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, (pp. 2555-2558). ACM. https://doi.org/10.1145/2396761.2398690

Kim, T., Barasz, K., & John, L. K. (2019). Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research*, *45*(5), 906-932. https://doi.org/10.1093/jcr/ucy039

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30-37. https://doi.org/10.1109/MC.2009.263

Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In *Recommender Systems Handbook* (pp. 77-118). Springer.

Krämer, J., & Schnurr, D. (2018). Is there a need for platform neutrality regulation in the EU?. *Telecommunications Policy*, *42*(7), 514-529. https://doi.org/10.1016/j.telpol.2018.06.004

Krämer, J., Schnurr, D., & Wohlfarth, M. (2019). Winners, losers, and facebook: The role of social logins in the online advertising ecosystem. *Management Science*, *65*(4), 1678-1699. https://doi.org/10.1287/mnsc.2017.3012

Krämer, J., Schnurr, D., & Wohlfarth, M. (2019). Trapped in the Data-Sharing Dilemma. *MIT Sloan Management Review*, *60*(2), 22-23.

Krämer, J., Senellart, P., and de Streel, A. (2020). *Making data portability more effective for the digital economy: Economic implications and regulatory challenges of the portability of personal data in the digital economy*. CERRE Policy Report. Centre on Regulation in Europe (CERRE).

Krämer, J., & Wohlfarth, M. (2018). Market power, regulatory convergence, and the role of data in digital markets. *Telecommunications Policy*, *42*(2), 154-171. https://doi.org/10.1016/j.telpol.2017.10.004

Krämer, J. and Zierke, O. (2020). *Paying for prominence: The effect of sponsored rankings on the incentives to invest in the quality of free content on dominant online platforms* (Working Paper). https://dx.doi.org/10.2139/ssrn.3584371

Krishnan, U., Moffat, A., Zobel, J., & Billerbeck, B. (2020, April). Generation of Synthetic Query Auto Completion Logs. In *European Conference on Information Retrieval* (pp. 621-635). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-45439-5_41

Kumar, A., & Hosanagar, K. (2019). Measuring the value of recommendation links on product demand. *Information Systems Research*, *30*(3), 819-838. https://doi.org/10.1287/isre.2018.0833

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, *13*(7), 759-780. https://doi.org/10.1080/17512786.2019.1567271

Lai, H. C., Shih, W. Y., Huang, J. L., & Chen, Y. C. (2016). Predicting traffic of online advertising in real-time bidding systems from perspective of demand-side platforms. In *2016 IEEE International Conference on Big Data (Big Data)*, (pp. 3491-3498). IEEE.

Lambrecht, A., & Tucker, C. E. (2015). Can Big Data protect a firm from competition?. *Available at SSRN 2705530*. http://dx.doi.org/10.2139/ssrn.2705530

Lee, D., & Hosanagar, K. (2016, April). When do recommender systems work the best? The moderating effects of product attributes and consumer reviews on recommender performance. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 85-97). https://doi.org/10.1145/2872427.2882976

Lee, D., & Hosanagar, K. (2018). How Do Product Attributes Moderate the Impact of Recommender Systems?. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2018/10/FP0315_WP_2018Oct.pdf

Lee, D., & Hosanagar, K. (2019). How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Information Systems Research*, *30*(1), 239-259. https://doi.org/10.1287/isre.2018.0800

Lerner, A.V. (2014). The Role of 'big data' in online platform competition. *Available at SSRN 2482780.* https://dx.doi.org/10.2139/ssrn.2482780

Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, *130*(4), 1941-1973. https://doi.org/10.1093/qje/qjv023

Li, B., Ch'ng, E., Chong, A. Y. L., & Bao, H. (2016). Predicting online e-marketplace sales performances: A big data approach. *Computers & Industrial Engineering*, *101*, 565-571. https://doi.org/10.1016/j.cie.2016.08.009

Li, X., Ling, C. X., & Wang, H. (2016). The convergence behavior of naive Bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(1), 1-24. https://doi.org/10.1145/2948068

Libert, T., Graves, L., & Nielsen, R. K. (2018). *Changes in third-party content on European news websites after GDPR.* Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-08/Changes%20in%20Third-Party%20Content%20on%20European%20News%20Websites%20after%20GDPR_0_0.pdf

Lin, Z., Goh, K. Y., & Heng, C. S. (2017). The Demand Effects of Product Recommendation Networks: An Empirical Analysis of Network Diversity and Stability. *MIS Quarterly*, *41*(2), 397-426. https://dx.doi.org/10.2139/ssrn.2389339

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76-80. https://doi.org/10.1109/MIC.2003.1167344

Linden, G. (2009, February 17). Jeff Dean keynote at WSDM 2009. *Geeking with Greg*. http://glinden.blogspot.com/2009/02/jeff-dean-keynote-at-wsdm-2009.html

Lindsey, N. (2019, November 25). *Google will restrict sharing of user data for Google Ads under EU privacy pressure*. CPO Magazine. https://www.cpomagazine.com/data-privacy/google-will-restrict-sharing-of-user-data-for-google-ads-under-eu-privacy-pressure/

Lomas, N. (2014, January 18). *Amazon patents "anticipatory" shipping – To start sending stuff before you've bought it.* Techcrunch. https://techcrunch.com/2014/01/18/amazon-pre-ships/?guccounter=1

Lomas, N. (2019, August 26). *Facebook succeeds in blocking German FCO's privacy-minded order against combining user data*. Techcrunch. https://techcrunch.com/2019/08/26/facebook-succeeds-in-blocking-german-fcos-privacy-minded-order-against-combining-user-data/

Ma, H., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011). Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 287-296). https://doi.org/10.1145/1935826.1935877

Macbeth, S. (2017). *Tracking the Trackers: Analysing the global tracking landscape with GhostRank*. Technical report, Ghostery, December 2017. https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf

Mandy, D. M., & Sappington, D. E. (2007). Incentives for sabotage in vertically related industries. *Journal of Regulatory Economics*, *31*(3), 235-260. https://doi.org/10.1007/s11149-006-9015-7

Martens, B. (2020). Data access, consumer interests, and social welfare: An economic perspective. *Available at SSRN 3605383*. https://dx.doi.org/10.2139/ssrn.3605383

Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining Massive Fine-Grained Behaviour Data to Improve Predictive Analytics. *MIS Quarterly*, *40*(4), 869-888.

Matthijs, N., & Radlinski, F. (2011, February). Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 25-34). ACM. https://doi.org/10.1145/1935826.1935840

Mattioli, D. (2020, April 23). *Amazon scooped up data from its own seller to launch competing products.* The Wall Street Journal. https://www.wsj.com/articles/amazon-scooped-up-data-from-its-own-sellers-to-launch-competing-products-11587650015

MacKenzie, I., Meyer, C., & Noble, S. (2013, October 1). *How retailers can keep up with consumers*. McKinsey. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

McAfee, P., J. Rao, A. Kannan, D. He, T. Qin, and T.-Y. Liu (2015). Measuring scale economics in search. http://www.learconference2015.com/wp-content/uploads/2014/11/McAfee-slides.pdf

McCarthy, T. (2017, May 26). *Amazon's first New York bookstore blends tradition with technology*. The Guardian. https://www.theguardian.com/technology/2017/may/26/amazon-new-york-bookstore

McLeod, J. (2020, February 6). *Inside the kill zone: Big Tech makes life miserable for some startups, but others embrace its power.* Financial Post. https://financialpost.com/technology/inside-the-kill-zone-big-tech-makes-life-miserable-for-some-startups-but-others-embrace-its-power

Mehta, N., Detroja, P., & Agashe, A. (2018, August 10). *Amazon changes prices on its products about every 10 minutes – here's how and why they do it*. Business Insider. https://www.businessinsider.de/international/amazon-price-changes-2018-8/?r=US&IR=T

Mellet, K., & Beauvisage, T. (2020). Cookie monsters. Anatomy of a digital market infrastructure. *Consumption Markets & Culture*, *23*(2), 110-129. https://doi.org/10.1080/10253866.2019.1661246

Metz, C. (2016, April 2). *AI is transforming Google Search. The Rest of the web is next.* Wired. https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/

Meyers, P. J. (2019, May 14). *How often does Google update its algorithm?*. Moz Blog. https://moz.com/blog/how-often-does-google-update-its-algorithm

Mitra, B., Diaz, F., & Craswell, N. (2017). Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, (pp. 1291-1299). International World Wide Web Conferences. https://doi.org/10.1145/3038912.3052579

Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, *13*(1), 1-126.

Monopolkommission. (2015). *Competition policy: The challenge of digital markets. Special Report 68.* https://www.monopolkommission.de/images/PDF/SG/s68_fulltext_eng.pdf

Moore, M. (2018). *Democracy hacked: Political turmoil and information welfare in the digital age.* Oneworld.

Motta, M. & Peitz, M. (2020). *Big Tech Mergers* (CEPR Discussion Paper no. 14353). Centre for Economic Policy Research (CEPR). https://cepr.org/content/free-dp-download-31-january-2020-competitive-effects-big-tech-mergers-and-implications

Mozilla. (2019, July 9). *Security/Anti tracking policy*. Mozilla Wiki. https://wiki.mozilla.org/Security/Anti_tracking_policy

Mueller, J. (2020, March 5). Announcing mobile first indexing for the whole web. *Google Webmaster Central Blog.* https://webmasters.googleblog.com/2020/03/announcing-mobile-first-indexing-for.html

Muller, B. (n.d.). *How search engines work: Crawling, indexing, and ranking*. MOZ. https://moz.com/beginners-guide-to-seo/how-search-engines-operate

Müller, O., Fay, M., & vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, *35*(2), 488-509. https://doi.org/10.1080/07421222.2018.1451955

Murgia, M. (2019, September 4). *Google accused of secretly feeding personal data to advertisers*. Financial Times. https://www.ft.com/content/e3e1697e-ce57-11e9-99a4-b5ded7a7fe3f

Murgia, M. (2020, March 16). *Google accused by rival of fundamental GDPR breaches.* Financial Times. https://www.ft.com/content/66dbc3ba-848a-4206-8b97-27c0e384ff27

Murphy, R. (2019a, December 11). *Local consumer review survey.* Bright Local. https://www.brightlocal.com/research/local-consumer-review-survey/

Murphy, R. (2019b, November 15). *Google Analytics for local business study*. Bright Local. https://www.brightlocal.com/research/google-analytics-for-local-businesses-study/

Nakashima, R. (2018, August 14). *AP exclusive: Google tracks your movement, like it or not*. AP News. https://apnews.com/828aefab64d4411bac257a07c1af0ecb

Napoli, P., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday, 22*(5). https://doi.org/10.5210/fm.v22i5.7051

Napoli, P. M. (2016). Special issue introduction: Big data and media management. *International Journal on Media Management, 18*(1), 1-7. https://doi.org/10.1080/14241277.2016.1185888

Neumann, N., Tucker, C. E., & Whitfield, T. (2019). Frontiers: How effective is third-party consumer profiling? Evidence from field studies. *Marketing Science*, *38*(6), 918-926. https://doi.org/10.1287/mksc.2019.1188

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges. *Queue*, *17*(2), 48-75.

Oberstein, M. (2019, April 10). What exactly is the difference between Google's neural matching & RankBrain. *RankRanger*. https://www.rankranger.com/blog/neural-matching-rankbrain-difference

O'conner, C. (2015, May 6). Announcing Google Cloud Bigtable: The same database that powers Google search, Gmail and Analytics is now available on Google Cloud Platform. *Google Cloud*

Platform Blog. https://cloudplatform.googleblog.com/2015/05/introducing-Google-Cloud-Bigtable.html

OECD. (2020). Line of Business Restrictions - Background note. OECD Working Party No 2 on Competition and Regulation. DAF/COMP/WP2(2020)1. https://one.oecd.org/document/DAF/COMP/WP2(2020)1/en/pdf
Oestreicher-Singer, G., & Sundararajan, A. (2012). The visible hand? Demand effects of recommendation networks in electronic markets. *Management Science*, *58*(11), 1963-1981. https://doi.org/10.1287/mnsc.1120.1536

OpenBanking. (2020). *Open Banking API performance*. https://www.openbanking.org.uk/providers/account-providers/api-performance/

Osmani, A., Gigorik, I. (2019, September 23). *Speed is now a landing page factor for Google Search ads.* Google. https://developers.google.com/web/updates/2018/07/search-ads-speed

Parboo, R. (2019, July 9). How to optimize your marketing attribution. *IAB UK.* https://www.iabuk.com/opinions/how-optimise-your-marketing-attribution

Parker, G., Petropoulos, G., & Van Alstyne, M. W. (2020). Digital Platforms and Antitrust. *Available at SSRN*. https://dx.doi.org/10.2139/ssrn.3608397

Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy? and How to Make Them Work for You*. WW Norton & Company.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159-188. https://doi.org/10.2753/MIS0742-1222270205

Pitkow, J. (2002). Personalized search: A content computer approach may prove a breakthrough in personalized search efficiency. *Communications of the ACM*, *45*(9), 50-55.

Porter, M. E., & Heppelmann, J. E. (2015). How smart, connected products are transforming companies. *Harvard Business Review*, *93*(10), 96-114.

Prager, A. (2019, May 10). *Vestager calls for more access to data for smaller platforms*. Euractiv. https://www.euractiv.com/section/data-protection/news/vestager-calls-for-more-access-to-data-for-small-platforms/

Preibusch, S., Peetz, T., Acar, G., & Berendt, B. (2016). Shopping for privacy: Purchase details leaked to PayPal. *Electronic Commerce Research and Applications*, *15*, 52-64. https://doi.org/10.1016/j.elerap.2015.11.004

Prüfer, J. (2020). *Competition Policy and Data Sharing on Data-driven Markets: Steps Towards Legal Implementation*. Friedrich Ebert Stiftung.

Prüfer, J. & Schottmüller, C. (2019). *Competing with Big Data* (TILEC Discussion Paper No. 2017-006). https://dx.doi.org/10.2139/ssrn.2918726

Ready, B. (2020, April 21). It's now free to sell on Google. *Google Blog*. https://blog.google/products/shopping/its-now-free-to-sell-on-google/

ReTV project. (n.d.). *ReTV-Project*. Retrieved from https://retv-project.eu/about/ on 2020, May 3

Rinehardt, W. (2018, November 7). *Is there a kill zone in tech?*. Tech Liberation. https://techliberation.com/2018/11/07/is-there-a-kill-zone-in-tech/

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, *10*(1), 1-9. https://www.nature.com/articles/s41467-019-10933-3

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2019). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181-1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Satariano, A. (2020, April 27). *Europe's privacy law hasn't shown its teeth, frustrating advocates*. The New York Times. https://www.nytimes.com/2020/04/27/technology/GDPR-privacy-law-europe.html

Schaefer, M., & Sapi, G. (2019). *Data Network Effects: The Example of Internet Search* (Working Paper). https://drive.google.com/file/d/1RRxhTW560PwtMGLEN-0wHikW7oVS9CEn/view?usp=sharing

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1-2), 115-153. https://doi.org/10.1023/A:1009804230409

Schepp, N. P., & Wambach, A. (2016). On big data and its relevance for market power assessment. *Journal of European Competition Law & Practice*, *7*(2), 120-124. https://doi.org/10.1093/jeclap/lpv091

Schiff, A. (2020, March 9). *Can LiveRamp survive the cookie apocalypse?*. Adexchanger. https://www.adexchanger.com/data-exchanges/can-liveramp-survive-the-cookie-apocalypse/

Schmidt, D. C. (2018). Google Data Collection. https://digitalcontentnext.org/wp-content/uploads/2018/08/DCN-Google-Data-Collection-Paper.pdf

Schumpeter, J. A. (1932). *The Theory of Economic Development: An inquiry into profits, capital, credit, interest, and the business cycle*. Harvard University Press.

Schwartz, B. (2016, October 18). *How Google uses machine learning in its search algorithm*. Search Engine Land. https://searchengineland.com/google-uses-machine-learning-search-algorithms-261158

Schwartz, B. (2019a, March 21). *Google's neural matching versus RankBrain: How Google uses each in search*. Search Engine Land. https://searchengineland.com/googles-neural-matching-versus-rankbrain-how-google-uses-each-in-search-314311

Schwartz, B. (2019b, December 17). *Google super easy explanation how it uses machine learning in search.* Search Engine Roundtable. https://www.seroundtable.com/google-explains-machine-learning-search-28697.html

Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Melamed, D. & Morgenstern, J. (2019). *Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report*. Draft. Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business. https://www.judiciary.senate.gov/imo/media/doc/market-structure-report%20-15-may-2019.pdf

ScrapeHero. (2019, April 24). *How many products does Amazon sell? – April 2019*. https://www.scrapehero.com/number-of-products-on-amazon-april-2019/

Search Engine News. (2020, July 1). *The unfair advantage book winning the search engine wars.* https://searchenginebook.com/

Sedhain, S., Menon, A. K., Sanner, S., & Xie, L. (2015). Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web* (pp. 111-112). https://doi.org/10.1145/2740908.2742726

Seeger, M. W., Salinas, D., & Flunkert, V. (2016). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* (pp. 4646-4654).

Segal, I., & Whinston, M. D. (2007). Antitrust in innovative industries. *American Economic Review*, *97*(5), 1703-1730. https://doi.org/10.1257/aer.97.5.1703

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, *80*(2), 159-169. https://doi.org/10.1016/j.jretai.2004.04.001

Shankland, S. (2014, November 19). *Firefox dumps Google for search, signs on with Yahoo*. Cnet. https://www.cnet.com/news/in-major-shift-firefox-to-use-yahoo-search-by-default-in-us/

Shapira, B., Rokach, L., & Freilikhman, S. (2013). Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction*, *23*(2-3), 211-247. https://doi.org/10.1007/s11257-012-9128-x

Shaw, D. (2018a, November 20). Announcing the 2018 Local Search ranking factors survey. *MOZ Blog*. https://moz.com/blog/2018-local-search-ranking-factors-survey

Shaw, D. (2018b, November 20). *2018 Local Search ranking factors.* MOZ. https://moz.com/local-search-ranking-factors

Singh, K., Vaver, J., Little, R. E., & Fan, R. (2018). *Attribution model evaluation.* Techreport, Google LLC.

Slawski, B. (2018, April 24). *PageRank update*. SEO by the Sea. https://www.seobythesea.com/2018/04/pagerank-updated/

Smith, B., & Linden, G. (2017). Two decades of recommender systems at amazon. com. *IEEE Internet Computing*, *21*(3), 12-18. https://doi.org/10.1109/MIC.2017.72

Smith, N. (2018, November 7). *Big Tech sets up a 'kill zone' for industry upstarts*. Bloomberg. https://www.bloomberg.com/opinion/articles/2018-11-07/big-tech-sets-up-a-kill-zone-for-industry-upstarts

Sokol, D. D., & Comerford, R. E. (2017). Does antitrust have a role to play in regulating big data? *Cambridge Handbook of Antitrust, Intellectual Property and High Tech*. Cambridge University Press.

Song, Y., Sahoo, N., & Ofek, E. (2019b). When and how to diversify - A multicategory utility model for personalized content recommendation. *Management Science*, *65*(8), 3737-3757. https://doi.org/10.1287/mnsc.2018.3127

Sørensen, J., & Kosta, S. (2019). Before and after GDPR: The changes in third party presence at public and private European websites. In *The World Wide Web Conference*, (pp. 1590-1600). ACM. https://doi.org/10.1145/3308558.3313524

Sterling, G. (2018, September 28). *Report: Google to pay Apple $9 billion to remain default search engine on Safari*. Search Engine Land. https://searchengineland.com/report-google-to-pay-apple-9-billion-to-remain-default-search-engine-on-safari-306082

Sullivan, D. (2016, June 23). *FAQ: All about the Google RankBrain algorithm.* Search Engine Land. https://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440

Tamine, L., & Daoud, M. (2018). Evaluation in contextual information retrieval: Foundations and recent advances within the challenges of context dynamicity and data privacy. *ACM Computing Surveys (CSUR)*, *51*(4), 1-36. https://doi.org/10.1145/3204940

Teevan, J., Dumais, S. T., & Horvitz, E. (2005, August). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 449-456). ACM. https://doi.org/10.1145/1076034.1076111

Think with Google. (2014, May). *Infographic: Understanding consumers' local search behavior.* https://www.thinkwithgoogle.com/advertising-channels/search/how-advertisers-can-extend-their-relevance-with-search-infographic/

Think with Google. (2016, May). *How mobile search connects consumers to stores*. https://www.thinkwithgoogle.com/consumer-insights/mobile-search-trends-consumers-to-stores/

Thirumalai, S., & Sinha, K. K. (2013). To personalize or not to personalize online purchase interactions: implications of self-selection by retailers. *Information Systems Research*, *24*(3), 683-708. https://doi.org/10.1287/isre.1120.0471

Tingleff, S. (2020, March 27). Explaining the privacy sandbox explainer. *IAB Tech Lab*. https://iabtechlab.com/blog/explaining-the-privacy-sandbox-explainers/

Toplensky, R. (2019, May 5). *Brussels poised to probe Apple over Spotify's fees complaint.* Financial Times. https://www.ft.com/content/1cc16026-6da7-11e9-80c7-60ee53e6681d

Toplensky, R. and Nicolaou, A. (2019, March 13). *Spotify files EU antitrust complaint against Apple*. Financial Times. https://www.ft.com/content/73e0d448-4577-11e9-a965-23d669740bfb

Toth, A. (2011, April 15). Updating our Log File Data Retention Policy to put data to work for consumers. *Yahoo! Policy Blog*. https://web.archive.org/web/20170224230903/http://www.ypolicyblog.com/policyblog/2011/04/15/updating-our-log-file-data-retention-policy-to-put-data-to-work-for-consumers/

Tucker, C. S., & Kim, H. M. (2009). Data-driven decision tree classification for product portfolio design optimization. *Journal of Computing and Information Science in Engineering*, *9*(4), 1-14. https://doi.org/10.1115/1.3243634

Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, *54*(4), 683-694. https://doi.org/10.1007/s11151-019-09693-7

Tung, L. (2017, November 15). *Google's back: It's Firefox default search engine again, after Mozilla ends Yahoo deal*. ZD Net. https://www.zdnet.com/article/googles-back-its-firefoxs-default-search-engine-again-after-mozilla-ends-yahoo-deal/

Turow, J. (2017). *The aisles have eyes: How retailers track your shopping, strip your privacy, and define your power*. Yale University Press.

Turow, J., & Couldry, N. (2018). Media as data extraction: Towards a new map of a transformed communications field. *Journal of Communication*, *68*(2), 415-423. https://doi.org/10.1093/joc/jqx011

Vinocur, N. (2019, December 27). *'We have a huge problem': European tech regulator despairs over lack of enforcement*. Politico. https://www.politico.com/news/2019/12/27/europe-gdpr-technology-regulation-089605

Wang, H., Wang, N., & Yeung, D. Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1235-1244).

Wang, W., & Benbasat, I. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*(3), 72-101. https://doi.org/10.17705/1jais.00065

Warren, E. (2019, March 8). *Here's how we can break up Big Tech*. Medium. https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97-121. https://doi.org/10.1509%2Fjm.15.0413

Wen, W., & Zhu, F. (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal*, *40*(9), 1336-1367. https://doi.org/10.1002/smj.3031

Wu, S., Ren, W., Yu, C., Chen, G., Zhang, D., & Zhu, J. (2016). Personal recommendation using deep recurrent neural networks in NetEase. In *2016 IEEE 32nd international conference on data engineering (ICDE)* (pp. 1218-1229). IEEE. https://doi.org/10.1109/ICDE.2016.7498326

Wu, T. (2018). *The curse of bigness: Antitrust in the new gilded age*. Columbia Global Reports.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403-414. https://doi.org/10.1002/bdm.2118

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science*, *66*(3), 1045-1070. https://doi.org/10.1287/mnsc.2018.3255

Yuan, H., Xu, W., & Wang, M. (2014, October). Can online user behavior improve the performance of sales prediction in E-commerce?. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2347-2352). IEEE. https://doi.org/10.1109/SMC.2014.6974277

Zamani, H., Bendersky, M., Wang, X., & Zhang, M. (2017, April). Situational context for ranking in personal search. In *Proceedings of the 26th International Conference on World Wide Web*, (pp. 1531-1540). International World Wide Web Conferences. https://doi.org/10.1145/3038912.3052648

Zhou, J., Albatal, R., & Gurrin, C. (2016). Applying visual user interest profiles for recommendation and personalisation. In *International Conference on Multimedia Modeling* (pp. 361-366). Springer.

Zhu, F., & Liu, Q. (2018). Competing with complementors: An empirical look at Amazon. com. *Strategic Management Journal*, *39*(10), 2618-2642. https://doi.org/10.1002/smj.2932

Zuckerberg, M. (2019, March 30). *Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas.* The Washington Post. https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

cerre

Centre on Regulation in Europe

Avenue Louise, 475 (box 10)
1050 Brussels, Belgium

+32 2 230 83 60

info@cerre.eu

cerre.eu

@CERRE_ThinkTank