How Important Are User-Generated Data for Search Result Quality?

Tobias J. Klein *Tilburg University*
Madina Kurmangaliyeva *Trinity College Dublin*
Jens Prüfer *Tilburg University*
Patricia Prüfer *Centerdata*

Abstract

Do search engines produce better results because their algorithms are better or because they can access more data from past searches? We document that the algorithm of a small search engine can produce nonpersonalized results that are of similar quality to those of the dominant firm (Google) for certain types of search queries. Overall differences in the quality of search results are explained by searches for rare queries, which constitute 74 percent of the traffic in our data. We conduct an experiment in which we keep the algorithm of a small search engine fixed and only vary the amount of data it uses as input. Our results show that giving small search engines access to more data about rare queries improves the quality of their results. This suggests that mandatory data sharing by large search engines is a necessary condition, yet probably not a sufficient one, to increase competition in the search market.

1. Introduction

Search engines are used by billions of people every day. They are an important part of the infrastructure for many other industries and are of very high economic, political, and social importance (Ducci 2020). Google is the dominant

Jens Prüfer is also affiliated with the University of East Anglia. We are extremely grateful to Marc Al-Hamez and Josep Pujol from Cliqz, who allowed us to run the experiment and answered many technical questions about the way search engines work. We are also grateful for the many useful comments we received when we presented the results of this paper on various occasions, including at the universities of Tilburg, Passau, Stockholm, Paris-Dauphine, East Anglia, and Toulouse (15th Digital Economics Conference) and at the virtual Economics of Platforms Seminar by the Toulouse School of Economics Digital Center, the 2022 Society for Institutional and Organizational Economics conference in Toronto, and the Tilburg Law and Economics Center annual conference. We would also like to thank Gregory Crawford for his discussion of our paper, Seyit Höcük and Pradeep Kumar from Centerdata for helping us to collect the data and preparing them for analysis, and Marco Alberti and Magdalena Kuyterink for excellent research assistance. We acknowledge funding from the German Ministry of Finance (grant fe 11/19). A replication package with data and code accompanies this essay.

[*Journal of Law and Economics,* vol. 68 (August 2025)]

provider of online searching, with a market share above 90 percent in many countries (Capala 2025). One potential explanation for this is that Google's algorithm generates search results that are better than the ones provided by its competitors. Another potential explanation is that Google has access to more relevant data and can therefore produce better search results (Halevy, Norvig, and Pereira 2009). Every time a user performs a search on Google, they click on some of the results; this generates data (user information) that are useful to produce better search results in the future.

Quantifying the contributions of these two explanations is highly relevant for policymaking, regulation, and firms' decision-making. If Google had only a better algorithm, then there would not be much reason for antitrust policy to intervene (Varian 2019; Bajari et al. 2019). Other firms could develop an algorithm that performs equally well and enter the market.[1] Google would find it worthwhile to invest in the algorithm to avoid this from happening. Or firms would invest in building an even better algorithm. In any of these cases, consumers would benefit from active competition. However, if Google was dominant mainly because it had access to more data, then this would mean that there was not a level playing field for Google and existing and potential competitors. It would be much more difficult for entering firms to provide search results that are of similar or better quality than Google's, even if Google would innovate very little. In such a situation, sharing user-generated data among competitors would help level the playing field between Google and (potential) competitors. This, in turn, might benefit consumers by increasing incentives for all search engines to innovate (Argenton and Prüfer 2012; Prüfer and Schottmüller 2021).

Users of search engines produce data by entering a search query (a search string submitted when they use the search engine) and selecting one result from the list the engine provides. Query logs record how many users select a given result for a given query. These data are useful input for providing search results. For two reasons, it is difficult to study empirically whether differences in search result quality are due to different algorithms or to different amounts of user information algorithms can use as input. The first reason is access to data. While search results are public, usage data are usually proprietary and cannot be accessed, which inhibits external review of empirical results by academic peers or public authorities (Persily and Tucker 2020). The second reason is that it is hard to identify the effect on search result quality caused by having access to more data related to past searches, which is the object of interest for policymaking (Lewis-Kraus 2022). This is hard because the number of past searches for the same or similar queries is not exogenous but likely correlated with the error term in an estimation equation for current searches. In this paper, we address both challenges.

"[A]n ideal experiment would be to fix the 'query difficulty' and exogenously provide more or less historical data" (He et al. 2017, p. 294). This paper reports the results of such an experiment. We collaborated with a small search engine,

[1] This is one way to interpret the introduction of large language models that are integrated into search engines (notably in Bing, which has integrated a version of ChatGPT since February 2023).

Cliqz (based in Munich, Germany).[2] It provided us with nonpersonalized search results for a representative set of queries for German users in April 2020. Importantly, it also provided us with a measure of the popularity of the underlying queries. Moreover, Cliqz conducted an experiment on our behalf in which it kept the search algorithm fixed and varied the amount of user-generated data used to produce search results. This allows us to conduct within-search-engine comparisons. In particular, it allows us to look at the search results for the same queries, varying only the amount of data used as input. We complement the Cliqz data with nonpersonalized search results from Google and Bing on the same queries in the same period in the same country. We asked external assessors to rate the quality of the search results on a Likert scale (but with no details on the origin of the results). This offers insights about between-search-engine comparisons.

We first report the results from an experiment in which we keep the search engine algorithm fixed and vary only the amount of data it uses as input. This produces causal evidence that having more user data on rare (or tail) queries enables search engines to produce higher-quality results. We find that, at the margin, having more data leads to a substantial increase in search result quality for rare queries and almost no increase for popular queries. Second, we compare quality across search engines and find that the algorithm of a small search engine can produce nonpersonalized results that are of similar quality to Google's. We do so by comparing the quality of search results for popular queries. By contrast, for rare search terms, our results show that Google's results quality depreciates only slightly relative to that of popular queries, whereas Cliqz's results quality decreases a lot. Finally, we compare average quality levels across search engines and find that our assessors evaluate Google's results above Bing's, which are evaluated above Cliqz's, which produces a ranking of perceived quality that is in line with the ranking of market shares in Germany.[3] Taken together, the findings in this paper show that the overall differences in the quality of search results are explained by rare queries. Crucial for competition policy, rare queries represent 74 percent of the traffic in our data. Hence, they represent the critical margin for competition.

Our paper contributes to the literature in at least two ways. It is well known that there is a positive relationship between the availability of past user data and search engine quality (He et al. 2017; Schaefer and Sapi 2023). Our paper is the first to quantify the importance of past user data for an actual entrant in the search engine industry. It shows that an entrant can produce nonpersonalized search results that are of similar quality to Google's for popular queries and identifies that rare queries constitute the critical challenge for competition in the search engine industry, where rare queries constitute the majority of all queries. This finding

[2] More details about the business background of Cliqz are in Online Appendix OA.

[3] Data are from Statcounter (https://gs.statcounter.com/search-engine-market-share/all/germany/).

is both of academic interest and highly policy relevant.[4] Indeed, lawmakers are currently in the process of preparing legislation to regulate dominant firms on platform markets (see Commission Regulation 2022/1926, On Contestable and Fair Markets in the Digital Sector and Amending Directives [EU] 2019/1937 and [EU] 2020/1828 [Digital Markets Act], 2022 O.J. [L 265] 1; UK Competition and Markets Authority 2020; American Innovation and Choice Online Act, H.R. 3816, 117th Cong., 2d sess. [2021]). The European Union's Digital Markets Act (DMA), which has been enforced since March 2024, prescribes that large "gatekeeper" firms must provide business users with data generated in the context of the use of their services (art. 6[10]). It even has a special clause on search engines, giving third-party providers of online search engines the right to ask gatekeepers for data on search queries that are generated by end users (art. 6[11]). Our results support this provision because they show that small search engines could most likely improve their results quality for rare queries—and hence increase the competitiveness of the market—if they had access to more user information.

Our second contribution is that we offer results from an experiment—one conducted by a small search engine that has not been previously studied. This allows us not only to shed some light on a provider that is usually not in the spotlight but also to obtain clean estimates of the dependence of search result quality on data as an input.

In Section 2, we provide a more detailed discussion of the literature to which our work relates and contributes. Section 3 provides details on the setup and the experiment. Section 4 reports the results. Section 5 discusses the results and concludes. This paper is deliberately short. The comprehensive Online Appendix contains many technical details and presents additional findings.

2. Background and Related Literature

Our paper seeks to contribute to the vast literature studying digital markets. Several papers provide experimental and simulation-based evidence related to the importance of data for platforms, for example, Sun et al. (2023), Wernerfelt et al. (2022), and Decarolis et al. (2020).[5]

Calvano and Polo (2021) conclude in their literature review that digital markets have a strong natural tendency toward concentration or market tipping, which suggests that models of competition for the market are more relevant than

[4] At least 30 top-level advisory reports about competition in online platform markets raise concerns related to markets in which data serve as an input (Beaton-Wells 2019), including highly regarded ones in the United States (Scott Morton et al. 2019), the European Union (Crémer, de Montjoye, and Schweitzer 2019), the United Kingdom (Furman et al. 2019), and Germany (Schallbruch, Schweitzer, and Wambach 2019).

[5] In addition, Microsoft argued that "obtaining the large quantity of data necessary to develop an effective [general] search engine (e.g., the information upon which relevancy algorithms can be built and improved) would be a significant barrier to entry" (European Commission, Commission Decision of 27.6.2017 relating to Proceedings under Article 102 of the Treaty on the Functioning of the European Union and Article 54 of the Agreement on the European Economic Area (AT.39740—Google Search (Shopping)), para. 286; brackets in original)

models for competition in the market. Krämer and Schnurr (2022) offer an excellent survey of the literature about economies of scale and scope in data and discuss various policy proposals, with a focus on data-sharing obligations.

Both the academic and the policy discussion about data sharing suffer from unclear definitions. Most of the literature studies situations in which a user knows more about their type or willingness to pay for a service than the provider of the service (see Bergemann, Bonatti, and Gan 2022 and the literature cited therein). Then, the voluntary balancing of that information makes markets more efficient (or enables follow-on innovation) but comes at a cost for the individual, including a decrease in privacy, and, hence, the net welfare effects may be positive or negative. By contrast, the search engine market is often referred to as a data-driven market, where the interaction between a service provider and a user is administered electronically such that it is possible to store users' choices (for example, clicking behavior) and characteristics (for example, location) with very little effort, that is, virtually for free. Hence, the one provider who interacts with a user already has access to the user's data at the start of the analysis. In such an environment mandatory data sharing is needed because one party, the incumbent, has no incentives to share voluntarily.

Prüfer and Schottmüller (2021, p. 967) define a market as data driven if a firm's marginal costs of innovation decrease in the amount of user information, that is, if it is subject to specific feedback effects ("data-driven indirect network effects"). They show in a dynamic model of competition in research and development that, in data-driven markets, user information leads to market tipping (monopolization) and low incentives to innovate both for the dominant firm and for (potential) challengers. The intuition is that the smaller firms, even if they are equipped with superior production technology, face higher marginal costs of innovation because they lack access to enough user information. If a smaller firm were to heavily invest in innovation and roll out its high-quality product, the dominant firm could imitate it quickly—at a lower cost of innovation—and regain its quality lead. Foreseeing this situation, rational entrepreneurs and private financiers would not invest in such a smaller firm in the first place.[6] The dominant firm knows about the deterring disincentive to innovate for its would-be competitors and can rest on a lower level of innovative efforts, too.

Some authors argue that mandating the sharing of (anonymized) data on user preferences and characteristics among competitors could mitigate market tipping and would have positive net effects on innovation and welfare if data-driven indirect network effects are sufficiently strong (Prüfer and Schottmüller 2021; Argenton and Prüfer 2012).

[6] This result is reflected by Edelman (2015, p. 397), who cites the oral testimony of Yelp's chief executive officer before the Senate Judiciary Subcommittee on Antitrust, Competition Policy, and Consumer Rights on September 21, 2011, and writes: "Google dulls the incentive to enter affected sectors. Leaders of TripAdvisor and Yelp, among others, report that they would not have started their companies had Google engaged in behaviors that later became commonplace."

Our paper aims to study how important user-generated data are for search result quality. We study this for nonpersonalized search results using across-search-engine comparisons and an experiment conducted with an actual entrant, Cliqz.[7] This complements the work by He et al. (2017) and Schaefer and Sapi (2023). He et al. (2017) study scale effects in web searches by comparing query logs of several hundred billion searches from two large search engines. They distinguish "popular queries, which do account for a majority of *searches*" from "rarer queries, which account for the majority of *queries*" (He et al. 2017, p. 295, emphasis in original). They measure search engine quality using the click-through rate (CTR), that is, the percentage of clicks on the top URL of a search result page. They document a concave relationship between the historical number of queries and the CTR.

Schaefer and Sapi (2023) use observational data from Yahoo.com to study whether there are economies of scale in Internet searching. They also focus on the CTR as a quality measure and show that having more data enhances search results quality and that having personal information (for instance, the ability of the search engine to track the browsing behavior of specific users) amplifies the speed of learning. Their findings are consistent with an incumbent data advantage due to the possession of personal information. A similar result is shown by Bajari et. al (2019) when studying Amazon data. They find that the prediction accuracy of their models increases with the time dimension (but with diminishing returns to scale).

Our results are broadly in line with those of He et al. (2017) and Schaefer and Sapi (2023) but go beyond the findings of these two papers. Since our study focuses on the quality of search results of an entrant, we can quantify the actual gap in search result quality between the entrant and the incumbent that is due to rare queries (and thus the availability of user-generated data).

Another interesting complementary paper is Lei, Chen, and Sen (2023), whose authors partnered with a search engine in China and ran a randomized control trial, also with nonpersonalized search results. In the control condition, they exposed users to the typical product of the smaller search engine, which was trained both by own (internal) data of the search engine and by external data, which they received via an application programming interface (API) from the market-leading Chinese search engine. In the treatment condition, they removed access to the API. They also show that the less data the algorithm has access to (in the treatment condition), the more search engine quality is reduced, as measured by the CTR. Confirming our results qualitatively, they conclude: "API removal leads to a 4.6% (±0.3%) decline in average CTR on search suggestions, highlighting the value of the market leader's data capability" (Lei, Chen, and Sen 2024, p. 4).

We also contribute to the existing empirical literature by using various measures of search engine quality: Unlike He et al. (2017), Schaefer and Sapi (2023), or Lei, Chen, and Sen (2024), we use human assessment and the overlap of Cliqz's

[7] See Hagiu and Wright (2023) for delineations and modeling of across-user learning and within-user learning.

search result sets with Google's as measures of quality. Finally, and most important, He et al. (2017) and Schaefer and Sapi (2023) use observational data, while we use an experimental setting to estimate the relationship between search result quality and the amount of data from past searches to which the search engine has access. Lei, Chen, and Sen (2024) also run an experiment but do so in the institutional context of China (while ours is Europe). Moreover, in their study, the target search engine's algorithm can (and has to) adjust to the consequences of having access to less data in the treatment condition, which changes various parameters. In our experiment, by contrast, we keep everything but access to data constant, and, hence, quality differences are directly related to the availability of data.

Finally, while search results from Google can be personalized to some degree,[8] a recent study based on German search data produced by Google finds that search results are highly regionalized but rarely personalized (Krafft et al. 2019). This finding increases the relevance of our results, as they are based on nonpersonalized data collected by a privacy-protective small search engine.

3. Data and Experimental Setup

We describe the data and experimental setup using the following conceptual framework. A search engine produces a set of results r for a given query q, a given web index I,[9] an algorithm with parameters θ, and data about previous searches in D.[10] That is,

$$r = f(q, I, D; \theta(D)).$$

Data in D have a direct effect on results (having more data is useful to predict what users are looking for) as well as an indirect effect (because they can be used to retrain and/or improve the algorithm). For our between-search-engine comparisons (details below), we evaluate the quality of search result set r for the same query q by three search engines operating with different indices I (WebFX Marketing Experts 2025), different sets of user data D, and different algorithms θ.[11] For our experiment (the within-search-engine comparison), we keep the function f, query q, index I, and algorithm θ constant and only vary the amount of data D to which the algorithm has access.

[8] See "Personalization and Search Results," Google Search Help, https://support.google.com/websearch/answer/12410098.

[9] A web index is a database used by search engines to store and organize information about web pages, which enables quick retrieval of relevant results when users perform searches.

[10] We think of r as a vector, q as a text string, I as a more complicated version of a table containing the list of web pages and information about them that can be collected, D as a table containing all the user-generated data the search engine has collected in the past (possibly aggregated in some way), and θ as a vector of parameters that are obtained in a first step for a given D, which is why we write $\theta(D)$. The term f is a vector-valued function.

[11] We think of the function f itself as the same across search engines, but this is not essential for the interpretation of the results, as one can think of the same f with a different θ as another function.

Table 1
Query Buckets

Bucket	Popular Queries by Threshold	Average Searches per Week	Implied Percentage of Traffic
1	.2	72.1	11
2	.8	9.8	6
3	4	3.2	10
4	20	1.2	18
5	75	1.0	56

Note. Bucket 1 contains the most popular queries and bucket 5 contains the least popular queries.

We collaborated with the small search engine Cliqz (see Online Appendix OA for background). The mission of Cliqz was to offer an alternative search engine that protects the privacy of its users. This is one reason we focus on nonpersonalized search results in this paper.

To construct the data we use in this paper, Cliqz ordered all search queries submitted by German users of its Human Web software from April 20 to April 26, 2020, by their frequency.[12] It formed five buckets in line with queries' frequency using the following thresholds: .2 percent, 1 percent, 5 percent, and 25 percent (so bucket 1 represents the top .2 percent of search queries by frequency, bucket 2 represents the next .8 percent of searches, and so on to add up to 100 percent). Table 1 shows that the distribution of traffic across queries is highly skewed; see also Goel et al. (2010). Bucket B5 contains the 75 percent least popular queries (tail queries), which generated 56 percent of the traffic.

We randomly drew 1,000 queries from each of the five buckets to construct a stratified sample of 5,000 queries. For each of those 5,000 queries, Cliqz gave us nonpersonalized search results at different levels of user information. Search results consist of a list of URLs and additional information on each item, like a short preview of the web page. The user data include so-called query log counts, which summarize how often URLs have been clicked on by past users for a given query. Importantly, the search algorithm detects the similarity between queries so that it can also use query log counts of similar queries to produce search results, for instance by applying cluster analysis to the query-URL bipartite graph (Liu et al. 2012; Sadikov et al. 2010). We asked Cliqz to keep the search engine's algorithm as it is and only vary the amount of user data used to produce search results on the night between April 26 and April 27. This is the experiment that we

[12] The Human Web is software integrated in the Cliqz browser or, alternatively, a software extension to Mozilla Firefox. It allows for the anonymous collection of users' browsing activity and user-generated query logs. For example, if a user of a Cliqz browser—or a Firefox browser with installed Cliqz extension—searched for "ebay auto" using Google, Bing, Cliqz, or any other search engine, the information on the search, the results, and the choices made by the user were transferred in an anonymized manner to Cliqz. Hence, these search queries represent all searches on any search engine for that subpopulation of users.

conducted. It helps us to provide direct evidence on the dependence of search results quality on the amount of data an algorithm uses.

Specifically, Cliqz provided results at 12 different levels of data on past searches: 100 percent (or full data), 90 percent, 80 percent, 70 percent, 60 percent, 50 percent, 40 percent, 30 percent, 20 percent, 10 percent, 1 percent, and .1 percent. To obtain respective query log counts, we multiply each query log count by the assumed fraction of available data and take the floor of that value as the new log count. For example, if a given query-URL pair has a count of 10 (that is, people who searched using that query clicked on that URL 10 times in the past), then the new count for that query-URL pair would be five for 50 percent of user data availability, one for 10 percent, and zero for 1 percent or below. Hence, if the Cliqz search engine had only 1 percent of its actual user data, it would completely miss that query-URL pair.

We complemented the Cliqz data with nonpersonalized search results from Google and Bing on the same 5,000 queries in the same period. For this, we used the API of a paid service for Google and Bing search engines. The API allowed us to specify that we wanted results for users from Germany, to make the results comparable with the search results of Cliqz.

Finally, we asked external assessors to rate the quality of the search results on a 7-point Likert scale for a subset of queries. We restrict attention to queries that are either in German or English and that are at least three characters long. Then, we sample 500 queries: We draw 50 queries from buckets 1 and 2 each, 100 queries from bucket 3, and 150 queries from each of buckets 4 and 5. We oversample rare queries (buckets 4 and 5) to reduce possible noise as we expect that rare queries might be more difficult to assess. We also remove seven queries with inappropriate content, which results in 493 queries for human assessment.

For each query, we keep seven result sets to assess: five for Cliqz (at 100 percent, 50 percent, 20 percent, 10 percent, and 1 percent of user data), one for Google, and one for Bing. We limit each result set to the top five results.[13] Additionally, for each sampled query, we construct a mixed top-five result set from those of Google, Bing, and Cliqz (at 100 percent of user data). We randomize the order in which Google, Bing, and Cliqz contribute to the mixed result set. See the details of the mixed result set construction in the Online Appendix.

The assessment design (more details of which are provided in the Online Appendix) ensures that the assessors did not know which search engine generated a set of search results. Hence, all brand effects are excluded. We hired two research assistants (RAs) at Tilburg University and 587 people in Germany (37 percent women, median age 34) through the Clickworker.com platform to perform the assessment (clickworkers). One RA received all the result sets corresponding to queries in German; the other received those for queries in English (each assessor was proficient in the relevant language). A total of 563 clickworkers each provided evaluations on average for 15 result sets. In total, each result set was evalu-

[13] The first five organic search results combined typically receive about 80 percent of the clicks for Google queries (see Chaffey 2024).

ated on average by four different people (one RA and three clickworkers). When evaluating the result sets, the assessors were able to click on the respective links. The Online Appendix contains further details, including the characteristics of the assessors and the instructions.

4. Results

4.1. The Experiment: Within-Search-Engine Comparisons

In the experiment, we start with a set of queries q, vary the query log counts in D_{Cliqz}, and keep I_{Cliqz} and θ_{Cliqz} constant. This allows us to obtain unambiguous results regarding the impact of having more data.[14] These results are conservative in two ways. First, in reality, the parameters θ_{Cliqz} would likely change when the algorithm was trained with less data and would produce worse search results. Second, we drop data proportionally (see Section 3). In reality, having less data would also generate noisier results.[15] For these two reasons, we would expect that the patterns we document would be stronger in reality, and in that sense our results are conservative.

Figure 1 shows that, for more popular queries (buckets 1–3), the search result quality increases with access to more data, but only until using about 20 percent of Cliqz's available data, which is already enough to produce the highest quality it can produce (as the curves start to flatten there). This supports statistical learning theory, which suggests diminishing returns to dataset size in terms of predictive performance (He et al. 2017; Varian 2019). The remaining differences may come from differences in the algorithm. Alternatively, they may be a consequence of complementary investments and organizational practices generating productivity gains from the use of data (Bresnahan, Brynjolfsson, and Hitt 2002; Bloom, Sadun, and Van Reenen 2012). Crucially, however, for rare queries (buckets 4–5) the quality of search results is at a much lower level, and the curves are still increasing when using 100 percent of data available to Cliqz. That is, even if the marginal increases in quality are decreasing, they are positive for all levels of data we could observe. This suggests that Cliqz's quality could benefit from access to more search log data on rare queries. However, whether Cliqz could eventually produce search results comparable to Google's remains an open question that our experiment cannot answer.

These results are based on between-subject comparisons. We pool assessments by the RAs and the clickworkers. In the Online Appendix, we show that these

[14] Rare queries may also differ from popular ones for other reasons. By conducting an experiment and holding the set of queries fixed, we control for these differences. For that reason in our analysis we do not have to control for any other query characteristics, for example, length or number of words or measures of the complexity of the query.

[15] To see this, suppose we start with a large number of query log counts, say 1,000 and 300. By the law of large numbers, the relative magnitudes are close to the ones for even more data. But the law of large numbers would fail to hold when we drop, for instance, 99 percent of the data. We instead act as if it still holds (and there is no noise from sampling error) by changing the query log counts to 10 and 3, respectively, with 1 percent of the original data.

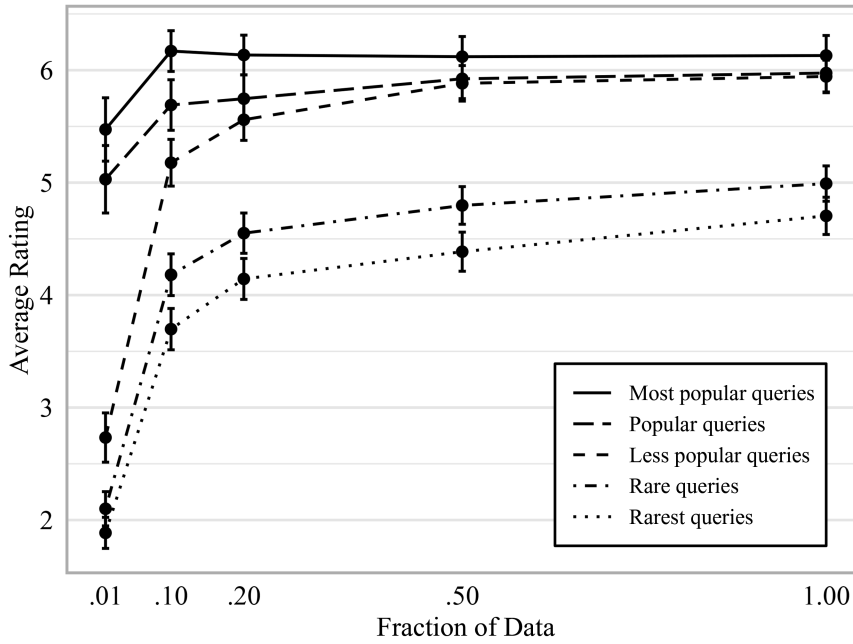

Figure 1. Estimates with 95 percent confidence intervals for the average quality of Cliqz's search results per bucket.

results are robust when we limit the sample to one type of assessor or when we measure only within-subject variation in the ratings. Even if there were scope for within-subject learning on what constitutes a good search engine result, because the order of result sets shown to the assessors was completely random, in expectation all buckets and all search engines should be affected equally. At the same time, we believe that both RAs and clickworkers were tech savvy enough to represent an average search engine user.

Next, we perform a robustness check. In Figure 2, we report how the overlap between Cliqz's and Google's top five results, which is an alternative measure of search results quality, depends on data available to Cliqz. The overlap is measured as the percentage of times the top five search results by Cliqz and Google are the same (in the sense that the set of results is the same without comparing the ordering). This implies that the independent variables in Figures 1 and 2 on the horizontal axis are the same but that the outcome variables are different (human-assessed quality versus overlap between Cliqz's and Google's results). Therefore, the magnitudes are not directly comparable. Yet, the qualitative findings are the same: For popular queries (buckets 1–3), about 20 percent of the data available to Cliqz's algorithm is sufficient to reach a level beyond which access to more data has minimal or no effects. For rare queries (buckets 4–5), however, there is no quality saturation. Having more data makes Cliqz's algorithm better in the

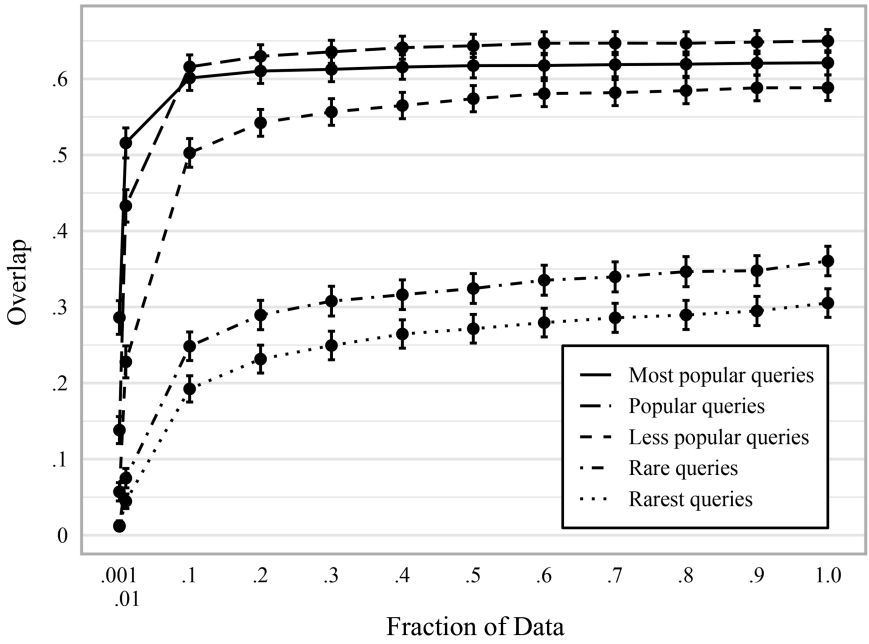

Figure 2. Estimates with 95 percent confidence intervals for the average quality of Cliqz's search results by the overlap with Google's results.

eyes of human assessors and brings its results closer to Google's, as measured by machine-calculated similarity scores.

Notably, in Figure 2 even for bucket 1, with the most popular queries, the overlap of Cliqz's results with Google's levels off at about 65 percent. As the human assessment has shown, this is not a sign of the diminished quality of Cliqz's results as compared with Google's: Users like the results for the most popular queries of both search engines equally (see Figure 2). However, there are various ways to answer certain queries, which is why only about three of the top five links produced by Cliqz (60 percent) are the same as those produced by Google.

4.2. Between-Search-Engine Comparisons

Figures 1 and 2 show that having more data on previous searches is an important ingredient in increasing search result quality, especially for rare queries. If such data are a driving factor of search result quality, as compared with other factors such as algorithmic quality, we should see differential access to data on previous searches reflected in the search result quality across search engines. By definition, a small search engine, which has fewer users than a large search engine with a high market share, should therefore be at a disadvantage in producing high-quality results, especially for rare queries.

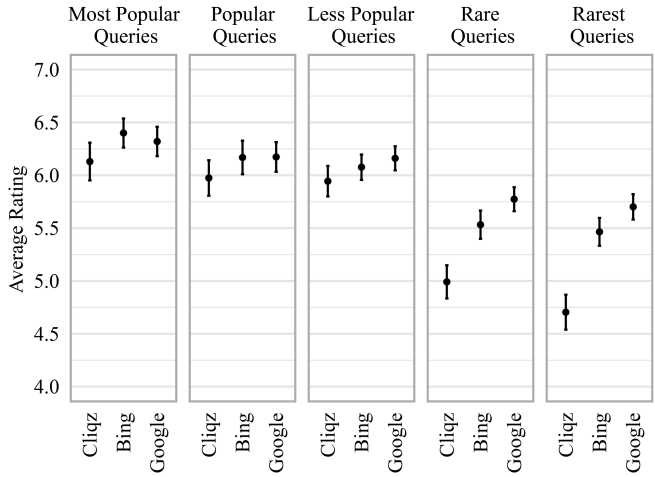

Figure 3. Estimates with 95 percent confidence intervals for the quality of search results by query popularity.

We also had assessors rate the search results of Google and Bing. This was done for the same queries, and the assessors did not know which search engine produced them. Figure 3 shows that for popular queries (buckets 1–3), search results are of comparable quality across search engines. This suggests that Cliqz's algorithm can provide similar quality to Google's for search queries when it can access substantial amounts of user information (see also Banko and Brill 2001). For less popular queries, however, Cliqz's results are significantly worse than Google's and Bing's.[16] These results reflect the differential role that data on previous searches play for popular versus rare queries: The quality of results from search engines with a higher market share and, hence, more data on both popular and rare queries decreases only a bit when moving from most popular to rarest queries; by contrast, a small search engine's quality of results depreciates significantly when comparing results for its most popular with its rarest queries.

This underlines that differences in the overall quality of search results are largely driven by differences in the quality of results for less popular queries. Table 1 shows that, on average, queries in bucket 4 and bucket 5 are searched only 1.2 and 1 times per week, respectively. It appears that this is not enough to produce sufficient data that can be used to produce high-quality search results. Crucially, buckets 4 and 5 jointly produce 74 percent of the traffic in our data: For a search engine to offer users satisfactory results, it is pivotal to perform well on those rare and rarest queries.

[16] Referring to our theoretical framework above, differences could also stem from different web indices I—Cliqz had indeed invested in its own web index. However, the quality of a web index is largely a function of money invested, just as the quality of an algorithm is. Economically they are the same in our production-function framework. Therefore, here we lump effects of the index I and the algorithm θ on results r together and distinguish both from data about previous searches D.

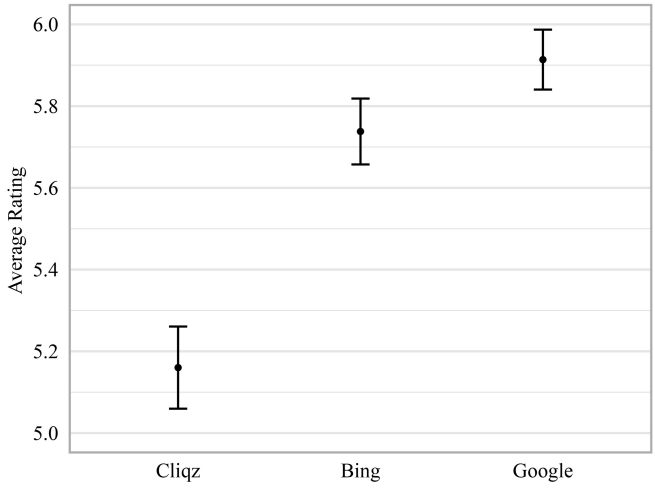

Figure 4. Estimates with 95 percent confidence intervals for the average quality of search results.

While this is a plausible explanation, in principle it could also be true that differences in the quality of search results for rare queries are driven by differences in the algorithm. To see this, let Google's algorithm have parameters θ_{Google}, and use data D_{Google} and index I_{Google}, while Cliqz's algorithm has parameters θ_{Cliqz}, and use data D_{Cliqz} and index I_{Cliqz}. Then, it could—again, in principle—be true that both search engines produce results that are of similar quality for popular queries and of different quality for rare queries solely because of differences between θ_{Google} and θ_{Cliqz}. This would be the case when Google's algorithm was particularly good at producing search results for rare queries. However, the within-search-engine experiment we reported on in Section 4.1 shows that data matters as an input and that at least some of the differences are related to the availability of data.

Finally, we bring both parts of the analysis together by comparing the overall average quality of search results across search engines. Figure 4 shows that the average quality of search results differs: Google produces the best results, followed by the number two in the market, Bing. The results produced by Cliqz are substantially worse. This ordering corresponds to the ordering of the market shares of the three search engines at the time of our experiment.

If we consider that search engine users' consumption choices, on top of ordinal quality considerations, are influenced by both legitimate factors (given that usage prices are nominal across all search engines, there is little reason to use any but the highest quality provider) and illegitimate factors (Google has been accused of various abuses of its monopoly/dominant position[17]), it is conceivable that the

[17] For instance, Google has been accused of placing its own products above competing sellers' if a user searches for products (European Commission, Commission Decision of 27.6.2017). Recently,

moderate differences between Google and Bing in Figure 4 are translated into drastic differences in market shares. Notably, the analysis presented here is orthogonal to these legal cases and to Google's conduct in markets. We show that having more data on rare queries makes search engines perform better and that the different access levels to such data across Google, Bing, and Cliqz in the market are consistent with our findings.

5. Discussion and Conclusion

Data are an important input for many services that are widely consumed (Mayer-Schönberger and Ramge 2018). In this paper, we study the importance of data about previous searches as an input in the context of online searching. We start from the empirical observation that Google has market shares around 90 percent in many markets, including the United States and Europe. To quantify the relevance of having more data about previous searches, compared with all other influences (including algorithmic quality), to the quality of results from today's search engines, we conduct an experiment with the small search engine Cliqz in which we vary only the amount of data Cliqz uses to produce search results. We show by two completely different outcome measures that data access has a significant effect on search result quality. This effect levels off for popular search terms, but a key result of this paper is that the quality returns to data are monotonically increasing for rare queries at the level of data Cliqz had during the time of our experiment.

Our within-search-engine results are reflected in our between-search-engine comparisons, which show that Cliqz's algorithm can produce results of similar quality to Google's and Bing's for popular search terms, that is, for queries when Cliqz has access to enough data. For rare queries, however, when Cliqz's algorithm can draw on data from only one previous search per week, search result quality decreases substantially, whereas the quality of Google's results and, to some extent, Bing's results decreases only a bit. Our final analysis brings all parts together and shows that the average quality ranking across search engines reproduces the ranking found in the market.[18]

Taken together, these findings show that search result quality depends on the amount of data a search engine has access to and that using less data has a particularly big effect on search result quality for rare queries. This suggests in turn that having access to more data would enable small search engines such as Cliqz to produce results for rare queries that are of better quality. This finding is highly policy relevant, as rare queries constitute most traffic (74 percent in our data).

a federal judge ruled that Google violated US antitrust law by maintaining a monopoly in the search and advertisement markets, especially by concluding exclusive deals with Apple and other operating system producers to install Google as the default search engine (Feiner 2024).

[18] Notably, shortly after our experiment was conducted, Cliqz announced that it would go out of business (see https://cliqz.com/announcement.html).

For that reason, entrants may have a hard time competing with incumbents.[19] A proposed remedy that might help level the playing field could be to require large search engines to share data about previous searches with smaller competitors. Unlike in other contexts, this remedy would not directly harm the incumbent because user data are nonrival. Data sharing only removes the incumbent's exclusivity advantage to access data.[20] Notably, saving raw data on previous searches is a by-product of operating a search engine that is independent of the fixed cost of operations, which can be huge. Moreover, big tech firms are already engaged in all kinds of voluntary data sharing, which suggests that even the costs to share large amounts of data are not prohibitive.[21] To the extent this also holds on the search engine market, data sharing can be a useful policy, too, specifically because data are nonrival.

Consequently, subjecting a dominant search engine, which has already sunk setup costs, to a data-sharing obligation would not disrupt its cost structure too much. On the other side, access to the shared data would improve the quality of competitors and newcomers, as shown in this paper, which would thereby make the search engine market more competitive. Consequently, these competitors would know they stand a chance in the market and have an incentive to innovate, either by improving their algorithms or by offering complementary services. This, in turn, would constitute stronger incentives for the dominant search engine to innovate, if it wants to stay in business, than remaining without mandatory data sharing (Prüfer and Schottmüller 2021). Consumers would benefit from both changes.[22]

Importantly, mandatory data sharing is not a silver bullet to reinstate level competition in the search engine market. Google, the dominant market leader, not only has a 90 percent market share or higher in many markets,[23] it also has

[19] One may wonder how, then, Google could have reached market leadership against Yahoo, the leading search engine in the 1990s. This development can be understood through the lens of the theory of data-driven markets (Prüfer and Schottmüller 2021): In the 1990s, Yahoo categorized websites manually. Google's original PageRank algorithm, as described in Page et al. (1998), let the rank of a website be determined only by the structure of hyperlinks on the World Wide Web, not by user information. Only since 2001 has Google made use of the data created by logging users' clicking behavior in search logs (Zuboff 2019). This innovation transformed the search engine industry into a data-driven market, leading to Google's market leadership as of 2003.

[20] See Graef and Prüfer (2021) for a fully fledged proposal of how to implement mandatory data sharing, including a governance structure that identifies who should be responsible for which tasks, that is in line with EU laws on privacy protection, intellectual property, consumer rights, and competition. See Krämer and Schnurr (2022) for a discussion of several ways of sharing data, including pros and cons of each.

[21] Recently, Google, Facebook, Microsoft, Twitter, and other firms showed that sharing user data is technically and organizationally possible at a large and automatic scale. They announced a new standards initiative called the Data Transfer Project, designed as a new way to move data between platforms (Brandom 2018).

[22] This is the case as long as the part of the additional fixed cost associated with operating multiple search engines that is passed on to consumers would not be too high. That cost could in principle be passed on indirectly in the form of more advertising that consumers are exposed to, even if using the search functionality or saving raw data on previous searches remains free, as it is now. At the same time, competition between search engines that results from leveling the playing field may discipline firms when it comes to passing such fixed costs on to consumers.

[23] Data are from Statcounter (https://gs.statcounter.com/search-engine-market-share).

dedicated hardware and data centers and a very strong brand name, and it can pay hardware manufacturers, such as Apple, to install it as the default search engine on millions of end-user devices.[24] However, the positive and monotonic role that having more data on past searches plays for rare queries, which constitute over 70 percent of all queries and hence strongly affect users' expectations about search engine quality, will draw them toward the search engine with the highest expected quality (see Figure 4). Therefore, it seems wise to provide small search engines with more data about rare queries independent of all other dimensions of competition. Another way of putting this is to say that data sharing is a necessary condition for competition on the search engine market but not a sufficient condition.

Finally, as a litmus test for the relative impact of data versus algorithmic quality, one could take the recent surge of large language models, notably ChatGPT, since November 2022. ChatGPT's integration into Bing since February 2023, following Microsoft's investment of US$10 billion in ChatGPT's owner, OpenAI, could be regarded as a very large investment in Bing's algorithmic quality. Several commentators expected that this would reshuffle market shares in the search engine market.[25] Nevertheless, empirically global market shares have been virtually fixed over the past 2 years.[26] Even the huge investment in Bing's algorithm has not put a dent in Google's 90 percent market share.

References

Argenton, Cédric, and Jens Prüfer. 2012. Search Engine Competition with Network Externalities. *Journal of Competition Law and Economics* 8:73–105.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. 2019. The Impact of Big Data on Firm Performance: An Empirical Investigation. *AER Papers and Proceedings* 109:33–37.

Banko, Michele, and Eric Brill. 2001. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073017.

Beaton-Wells, Caron. 2019. Ten Things to Know about the ACCC's Digital Platforms Inquiry. *CPI Oceania Column,* August. https://www.competitionpolicyinternational.com/wp-content/uploads/2019/08/Oceania-Column-August-2019-Full.pdf.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan. 2022. The Economics of Social Data. *RAND Journal of Economics* 53:263–96.

[24] In a survey we conducted with a representative sample of the Dutch population (in the so-called LISS panel), among 538 individuals who were aware of other search engines than Google the three most frequently mentioned reasons to use Google were that Google provides high-quality search results, is a well-known search engine, and is the standard search engine on an individual's most-used device.

[25] See Prüfer (2023) for some background and references about those discussions and a theory-led perspective on the likely impact of large language models on the search engine market, as of February 2023.

[26] See Statcounter with a time window set to May 2022–September 2024 (https://gs.statcounter.com/search-engine-market-share).

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2012. Americans Do IT Better: US Multinationals and the Productivity Miracle. *American Economic Review* 102:167–201.

Brandom, Russell. 2018. Google, Facebook, Microsoft, and Twitter Partner for Ambitious New Data Project. *Verge,* July 20. https://www.theverge.com/2018/7/20/17589246/data-transfer-projectgoogle-facebook-microsoft-twitter.

Bresnahan, Timothy F., Erik Brynjolfsson, and Loren M. Hitt. 2002. Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *Quarterly Journal of Economics* 117:339–76.

Calvano, Emilio, and Michele Polo. 2021. Market Power, Competition, and Innovation in Digital Markets: A Survey. *Information Economics and Policy* 54, art. 100853, pp. 1–18.

Capala, Matthew. 2025. Global Search Engine Market Share in the Top 15 GDP Nations (Updated for 2025). Alphametric, January 10. https://alphametic.com/global-search-engine-market-share.

Chaffey, Dave. 2024. 2024 Comparison of Google Organic Clickthrough Rates (SEO CTR) by Ranking Position. Smart Insights, January 19. https://www.smartinsights.com/search-engine-optimisation-seo/seo-analytics/comparison-of-google-clickthrough-rates-by-position/.

Crémer, Jacques, Yves-Alexandre de Montjoye, and Heike Schweitzer. 2019. *Competition Policy for the Digital Era.* Publications Office of the European Union. https://data.europa.eu/doi/10.2763/407537.

Decarolis, Francesco, Gabriele Rovigatti, Michele Rovigatti, and Ksenia Shakhgildyan. 2020. Artificial Intelligence and Data Obfuscation: Algorithmic Competition in Digital Ad Auctions. Discussion Paper No. 18009. Centre for Economic Policy Research, London.

Ducci, Francesco. 2020. *Natural Monopolies in Digital Platform Markets.* Cambridge University Press.

Edelman, Benjamin. 2015. Does Google Leverage Market Power through Tying and Bundling? *Journal of Competition Law and Economics* 11:365–400.

Feiner, Laura. 2024. Judge Rules That Google "Is a Monopolist" in US Antitrust Case. *Verge,* August 5. https://www.theverge.com/2024/8/5/24155520/judge-rules-on-us-doj-v-google-antitrust-search-suit.

Furman, Jason, Diane Coyle, Amelia Fletcher, Philip Marsden, and Derek McAuley. 2019. *Unlocking Digital Competition: Report of the Digital Competition Expert Panel.* UK Treasury.

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. Predicting Consumer Behavior with Web Search. *PNAS* 107:17486–90.

Graef, Inge, and Jens Prüfer. 2021. Governance of Data Sharing: A Law and Economics Proposal. *Research Policy* 50, art. 104330, pp. 1–14.

Hagiu, Andrei, and Julian Wright. 2023. Data-Enabled Learning, Network Effects, and Competitive Advantage. *RAND Journal of Economics* 54:638–67.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24:8–12.

He, Di, Aadharsh Kannan, Tie-Yan Liu, R. Preston McAfee, Tao Qin, and Justin M. Rao. 2017. Scale Effects in Web Search. *Web and Internet Economics: 13th International Conference, WINE 2017,* edited by Nikhil R. Devanur and Pinyan Lu. Springer. https://doi.org/10.1007/978-3-319-71924-5_21.

Krafft, Tobias D., Michael Gamer, and Katharina A. Zweig. 2019. What Did You See? A Study to Measure Personalization in Google's Search Engine. *EPJ Data Science* 8, art. 38. https://doi.org/10.1140/epjds/s13688-019-0217-5.

Krämer, Jan, and Daniel Schnurr. 2022. Big Data and Digital Markets Contestability: Theory of Harm and Data Access Remedies. *Journal of Competition Law and Economics* 18:255–322. https://doi.org/10.1093/joclec/nhab015.

Lei, Xiaoxia, Yixing Chen, and Ananya Sen. 2024. The Value of External Data Capabilities in the Search Market: Evidence from a Field Experiment. Working paper. Shanghai Jiao Tong University, Shanghai.

Lewis-Kraus, Gideon. 2022. How Harmful Is Social Media? *New Yorker,* June 3. https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think.

Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic Taxonomy Construction from Keywords. In *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery.

Mayer-Schönberger, Viktor, and Thomas Ramge. 2018. *Reinventing Capitalism in the Age of Big Data.* Basic Books.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford University InfoLab, Stanford.

Persily, Nathaniel, and Joshua A. Tucker. 2020. *Social Media and Democracy: The State of the Field and Prospects for Reform*. Cambridge University Press.

Prüfer, Jens. 2023. February 8. ChatGPT vs. Bard in Search Engines: Good to Have a Theory. Prufer.net (blog), February 8. https://prufer.net/2023/02/08/chatgpt-vs-bard-in-search-engines-good-to-have-a-theory/.

Prüfer, Jens, and Christoph Schottmüller. 2021. Competing with Big Data. *Journal of Industrial Economics* 69:967–1008.

Sadikov, Eldar, Jayant Madhavan, Lu Wang, and Alon Halevy. 2010. Clustering Query Refinements by User Intent. In *WWW 2010: Proceedings of the 19th International Conference on World Wide Web*. Association for Computing Machinery.

Schaefer, Maximilian, and Geza Sapi. 2023. Complementarities in Learning from Data: Insights from General Search. *Information Economics and Policy* 65, art. 101063, pp. 1–19.

Schallbruch, Martin, Heike Schweitzer, and Achim Wambach. 2019. *Ein neuer Wettbewerbsrahmen für die Digitalwirtschaft*. Federal Ministry for Economic Affairs and Energy Public Relations Department.

Scott Morton, Fiona, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A. Douglas Melamed, and Jamie Morgenstern. 2019. *Market Structure and Antitrust Subcommittee Report*. University of Chicago Stigler Center for the Study of the Economy and the State, Chicago. https://www.chicagobooth.edu/-/media/research/stigler/pdfs/market-structure-report.pdf.

Sun, Tianshu, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu. 2023. The Value of Personal Data in Internet Commerce: A High-Stakes Field Experiment on Data Regulation Policy. *Management Science* 70:2645–60.

UK Competition and Markets Authority. 2020. *Online Platforms and Digital Advertising.* UK Competition and Markets Authority. https://assets.publishing.service.gov.uk/media/5efc57ed3a6f4023d242ed56/Final_report_1_July_2020_.pdf.

Varian, Hal. 2019. Artificial Intelligence, Economics, and Industrial Organization. Pp. 399–419 in *The Economics of Artificial Intelligence: An Agenda,* edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press.

WebFX Marketing Experts. 2025. How Search Engines Work: Crawling, Indexing, Ranking, and More. SEO.com, May 6. https://www.seo.com/basics/how-search-engines-work/.

Wernerfelt, Nils, Anna Tuchman, Bradley Shapiro, and Robert Moakler. 2022. Estimating the Value of Offsite Data to Advertisers on Meta. Working Paper No. 114. University of Chicago Becker Friedman Institute for Economics, Chicago.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.