The Value of External Data for Digital Platforms: Evidence from a Field Experiment on Search Suggestions*

Xiaoxia Lei
Shanghai Jiao Tong University†

Yixing Chen
University of Notre Dame‡

Ananya Sen
Carnegie Mellon University§

February 2023

Abstract

Firms increasingly leverage external data with an aim to unlock improvements in products and services, but it is challenging to measure the value of external data. Collaborating with a large Chinese technology company, we analyze a randomized field experiment where we manipulate access to the market leader's application programming interface (API) to measure the causal impact of depersonalized external data on click-through rate (CTR) for the focal company's nascent search product. We report three main findings: First, compared to the baseline with access to the market leader's API, API removal leads to a 4.6% decrease in CTR for search suggestions. Second, the negative effect due to API removal is more prevalent among heavy users, when users' search intent is implicit, and for mainstream (popular) content. Third, the magnitude of this negative effect in the long run is half as much as what we would have obtained with a short-term experiment. This research informs managers of whether and how the market leader's data affects a smaller player's product performance. It also sheds light on policies such as the Digital Markets Act that proposes privacy preserving data sharing by large digital platforms as well as a recent debate on whether big data undermines market competition.

*We thank John Lalor, Alex Moehring, Christian Peukert, Yoonseock Son, Joy Wu, and workshop and seminar participants at University of Notre Dame (Marketing), Carnegie Mellon University (Tepper School of Business), NYU AI in Strategic Management Workshop, NSF Convergence Workshop on Human-AI Frontier, China India Insights Conference, Conference on Information Systems and Technology, Marketing Dynamics Conference, and Workshop on Information Systems and Economics for helpful suggestions.

†Antai College of Economics and Management, Shanghai Jiao Tong University, xiaoxia.lei@sjtu.edu.cn.

‡Mendoza College of Business, University of Notre Dame, ychen43@nd.edu.

§Heinz College, Carnegie Mellon University, ananyase@andrew.cmu.edu.

1 Introduction

Data is considered the 'new oil' for digital platforms (Economist, 2017). Digital platforms collect large amounts of data such as their users' digital footprints and use it as a key input to develop algorithms that provide personalized recommendations (Sun et al., 2021). Increasingly, firms have tapped into 'external data', often through data-sharing agreements, with an aim to unlock improvements in products and services.[1] A prevalent practice is that firms use Application Programming Interfaces (APIs) to gain access to external data from large players in the marketplace (Benzell et al., 2022; Xue et al., 2019). For example, Google custom search JSON API enables publishers and developers to retrieve response data (e.g., search results) from Google. Yet there is limited evidence on the value of such external data for these recipient firms because it is challenging to pin down its causal impact on their product performance. Measuring the causal impact of external data is important not only from a managerial perspective but also from a policy perspective. The Digital Markets Act that came into effect in November 2022 requires platforms that provide core services (e.g., search engines, social networking services) to open up for smaller players with data access. In particular, the Digital Markets Act mandates large players in the search market to share *depersonalized* search results to small players and new entrants to level the playing field.[2]

In this paper, we analyze a large-scale field experiment to measure the extent to which *depersonalized* external data shared by a leading search engine (i.e., the market leader) affects search product performance of our partner company (i.e., a smaller player; hereafter, "the company"). We further explore *which* and *how* users respond to such data in the short term as well as how user behavior evolves in the longer term in response to the experimental variation. The primary outcome of interest is click-through rate (CTR), a key performance indicator of search success, online advertising effectiveness, and online behavior at large. The company made its initial foray into the search market by launching a new search product in 2020: i.e., search suggestions. Similar to the Google autocomplete predictions,[3] search suggestions are generated by the underlying

[1] See the survey report by Deloitte here https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html.

[2] See details about the proposal of for a regulation on DMA:https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en#documents.

[3] https://blog.google/products/search/how-google-autocomplete-works-search/.

algorithm that uses internal, personalized and external, depersonalized data, to predict what users want to click in response to their queries. Providing users with relevant suggestions through query autocomplete is fundamental to user experience, so large players devote significant resources to the development of search suggestions (Agrawal et al., 2018). Since the company was an entrant, it started by leveraging the market leader's data to improve search suggestions. We partnered with the search product team to experiment with access to the market leader's API while using the same algorithm to generate search suggestions, providing an ideal context to measure the value of external data.

The field experiment lasted 108 days and involved more than 2.3 million users among which each user was randomly assigned to one of two conditions. In the control condition (the status quo), the company's ranking algorithm ranks depersonalized search-suggestion candidates retrieved from the market leader's API along with its own search-suggestion candidates and generates a final list of search suggestions in response to user-submitted queries. In the treatment condition, we remove access to the market leader's API: i.e., the company's ranking algorithm does not have access to the market leader's search-suggestion candidates and only ranks its own search suggestions candidates. This research design has two notable features. First, the manipulation is changing the supply of candidates at the ranking stage of search-suggestion generating process while holding the ranking algorithm constant. Therefore, the random assignment among millions of users enables us to precisely estimate the causal impact of access to the market leader's data. Second, maintaining a consistent treatment for about 16 weeks enables us to estimate longer-term effects of the market leader's data on user behavior as well.

We report three sets of findings. First, relative to the status quo, the removal of the market leader's API leads to a 4.6% ($\pm$0.3%) decline in CTR, highlighting the value relevance of the market leader's data to the company. We benchmark the magnitude of the average treatment effect with the extant literature and industry reports and find it economically meaningful. Our (back of the envelope) revenue computations suggest that such data can have a significant payoff to the company.

Second, we document who are more (less) responsive to the removal of the market leader's data and the conditions under which the market leader's data impacts users. The negative impact of

removing the market leader's data is stronger for heavy users relative to newer users, suggesting that such data tends to help in user retention by streamlining active users' experience rather than ameliorate the cold-start problem for newer users. In addition, API removal reduces the number of content categories searched, suggesting that the market leader's data expands the breadth of content. Next, the reduction in CTR is primarily driven by relatively popular (mainstream) topics, suggesting that access to the market leader's API helps identify "easy hits" as opposed to niche/long-tail content. Together, these results seem to suggest that the depersonalized nature of external data lends itself to better identify what the crowd prefers to match users' search intent. Finally, the decrease in CTR due to API removal is higher when users' search intent lacks clarity or is implicit, suggesting that the use of crowd preferences is less important when search intent is explicit and can be matched easily to a suggestion.

Third, the magnitude of treatment effect estimates varies significantly within the 16-week experimental period. It starts at about a 8.1%–9% decline in the first 2–3 weeks, decreasing to 3.6%–4.5% in the last 3–4 weeks. These estimates suggest that had we run a short-term experiment, we could have overestimated the value of the market leader's data (or the negative effect of API removal) by a factor of 2. Furthermore, this data pattern is neither driven by differential attrition across two conditions because we do not find a significant effect of API removal on login activities, nor by relative improvement in algorithmic recommendations because of similar trends in CTR over the course of the experiment. Finally, we uncover that API removal increases the demand for search button, demonstrating a positive demand spillover to search button as a substitute for search suggestions.

Together, our findings offer clear practical implications. First, the causal estimates inform managers of the potential value proposition when tapping into different external data levers to enhance product performance. For example, our analysis provides empirical evidence that relates to Google's recent decision on restricting smaller players' access to its Autocomplete API.[4] Second, our findings inform policy makers in particular with regards to the Digital Markets Act. Our research provides novel causal evidence that a gatekeeper's data does enhance a smaller player's product performance in the context of organic search. Our context is important from a policy

[4]https://developers.google.com/search/blog/2015/07/update-on-autocomplete-api.

perspective since regulation has focused on the need for large online platforms/gatekeepers to share depersonalized search results with newer competitors. More generally, we provide empirical evidence that informs a recent debate on whether big data undermines market competition (Crémer et al., 2019; Tucker, 2019). Third, our results show that depersonalized data can help a search product *only* through identifying more mainstream topics but not niche content, suggesting its limited ability in catering to individual preferences. Building on the empirical evidence on the value of personalization (Sun et al., 2021; Yoganarasimhan, 2020), we demonstrate that despite their privacy-preserving intent, alternative approaches such as depersonalized data-sharing mechanisms have limitations in unleashing the full potential of data.

Our paper contributes to three strands of academic literature. First, we contribute to the literature that aims to quantify the value (of different dimensions) of data for search engines and related online products. Yoganarasimhan (2020) analyzes how utilization of user-level data can help in personalization of search and quantifies significant returns to personal data. Chiou and Tucker (2017) analyze the efficacy of search recommendations when companies such as Yahoo! and Microsoft reduced the amount of individual-level data retention to a 90-day period and find no change in a user's CTR. Valavi et al. (2020) quantifies the dynamic value of data by using platform data over different time periods in a next word prediction task, similar to our autocomplete setting, with Reddit as their context. We contribute to this strand of research by showing how external data sources impact online product performance using a large-scale, longer-term field experiment.

We also contribute to a strand of literature that analyzes the impact of utilizing several forms of 'external data' on firm outcomes. These external sources can be varied. Beraja et al. (2020) show that access to government data, in the form of surveillance videos, lead to AI innovation in the facial recognition industry by private companies in China. Similarly, Nagaraj (2022) analyzes the private impact of public data in the Gold industry. Nagaraj (2022) finds that access to publicly available Landsat, a U.S. National Aeronautics and Space Administration satellite mapping program, lead to more gold discoveries especially by new entrants. In the case of online platforms, Neumann et al. (2019) highlight how third-party data from data brokers has become pervasive for online ad targeting and analyze the quality of such widespread third-party data. Given our experimental setup, we can

cleanly identify effects along with heterogeneity within a policy-relevant search context. Moreover, we focus on a widely used yet under-explored source of external data for online platforms, providing implications for other companies that aim to utilize such data for their product growth.

Finally, more broadly, we contribute to the literature that analyzes different strategies for platform growth. Benzell et al. (2022) use aggregate data on online platforms to demonstrate how a firm can grow by opening itself up to third party complementors using APIs. Sun et al. (2021) simulate a privacy regulation through a field experiment on Alibaba to quantify the value of individual-level data for users of the platform. They find significant negative effects on engagement and purchase when personal data is not used for product recommendations especially for niche products. Relatedly, Claussen et al. (2019) quantify the value of personal data (or lack thereof) for an algorithm relative to human experts. Relative to these papers, we take a middle path and analyze a privacy-preserving situation where depersonalized data from other (external) sources could be used as an input for algorithmic recommendations. We show that it can help engagement but only for mainstream content, demonstrating the trade-off in utilizing such data.

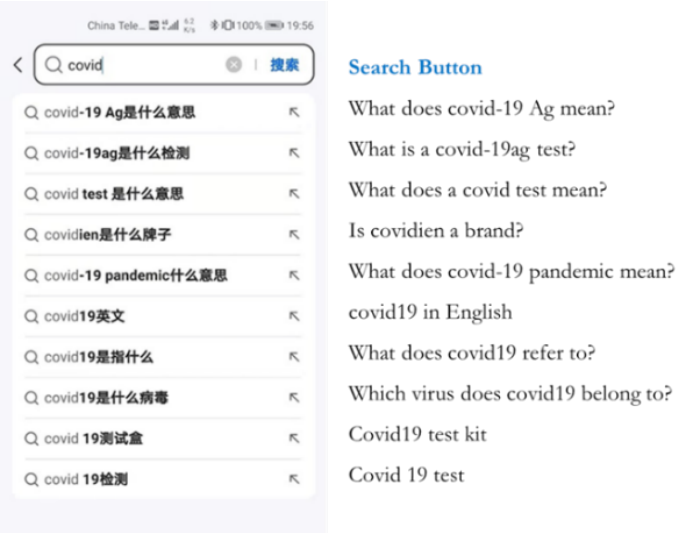
2 Empirical Setting

In this section, we describe the empirical setting with a focus on (1) search-suggestion generating process through the lens of the 'algorithmic funnel', (2) the role of the market leader's API in such a process, and (3) our choice of target metric that measures the performance of search suggestions.

We partnered with a large Chinese technology company that develops a web browser app. With more than 200 million monthly active users, the app offers its users various in-app products such as news feed, search engine, and video and eBooks streaming. The main search product is search suggestions, which was first launched in 2020. By design, search suggestions are predictive in nature and function as query autocomplete, which is similar to Google autocomplete feature offered by Google Chrome. Figure 1 provides an illustration of the app interface. When users enter a specific query term into the search bar (e.g., covid), they typically see a list of ten search suggestions in the form of phrases and/or sentences in response to the query (e.g., "What does a covid-19 Ag mean?").

A user can choose to click a matched suggestion, press the search button without adopting search suggestions, delete this query and enter a new one, or quit the session. Search suggestions and related contexts of predictive text is economically important with companies investing significant resources (Agrawal et al., 2018). Search suggestions bridge the gap between users' search intent and content consumption helping streamline the process. Recent evidence shows that such suggestions can influence decisions of individuals in financial markets (Rubin and Rubin, 2021). Optimizing search suggestions is important from the perspective of our partner company because they intend to monetize this process. In particular, their aim is to earn revenue from each click that is made on search suggestions in the same way sponsored links generate revenue for a search engine.

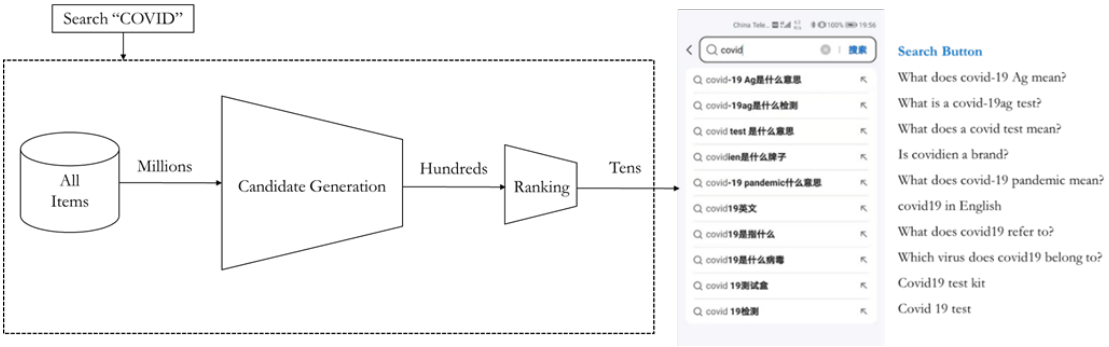
Figure 1: An Illustration of Search Suggestions



How does the company generate search suggestions in response to user-submitted queries? At a high level, the company uses a proprietary algorithmic funnel, a three-stage architecture similar to a recommendation system (Covington et al., 2016): (1) item generation, (2) candidate generation, and (3) ranking. Figure 2 provides an illustration. At the *item generation* stage, the company first builds raw items based on (a) internal sources in the news feed (e.g., articles and videos) and query terms from users' active search histories and (b) external sources such as public trending news. Next, the company uses natural language processing techniques (e.g., keyword extraction,

text summarization, and knowledge graph) to transform raw items into phrases and sentences. As a result, this large item base contains millions of phrases and sentences. At the *candidate generation* stage, a subset of candidate items are retrieved from the item base based on (a) their relevance to the original query term entered by a user and (b) their popularity based on real searches by all users on the platform (e.g., common and trending searches). As a result, hundreds of candidate phrases and sentences are selected to enter the ranking stage. At the *ranking* stage, a ranking algorithm scores each candidate item according to its predicted click-through rate, which is a function of user (e.g., location, search histories) and query (e.g., topic, freshness) features. The highest scoring items are presented in a ranked order in the final list to users. The literature shows that sophisticated algorithms are utilized to handle a larger feature set at the ranking stage and rank fewer items. This is in contrast to the candidate generation stage that uses basic algorithms and considers a much larger set of items focusing on pruning the data of duplicate and irrelevant items in an efficient manner (Covington et al., 2016; Nandy et al., 2021).

Figure 2: An Overview of the Algorithmic Funnel



However, a key challenge of developing a new search product such as search suggestions is the lack of candidate items with respect to both quantity and quality. For example, at the initial development stage, the company's algorithmic funnel can generate only a short list of search suggestions that could match users' search intent. To address this problem, the company had started to leverage proprietary data via access to the API of a leading search engine in China (i.e., the market leader). The company's hypothesis is that because of its ability to present the most relevant, useful results

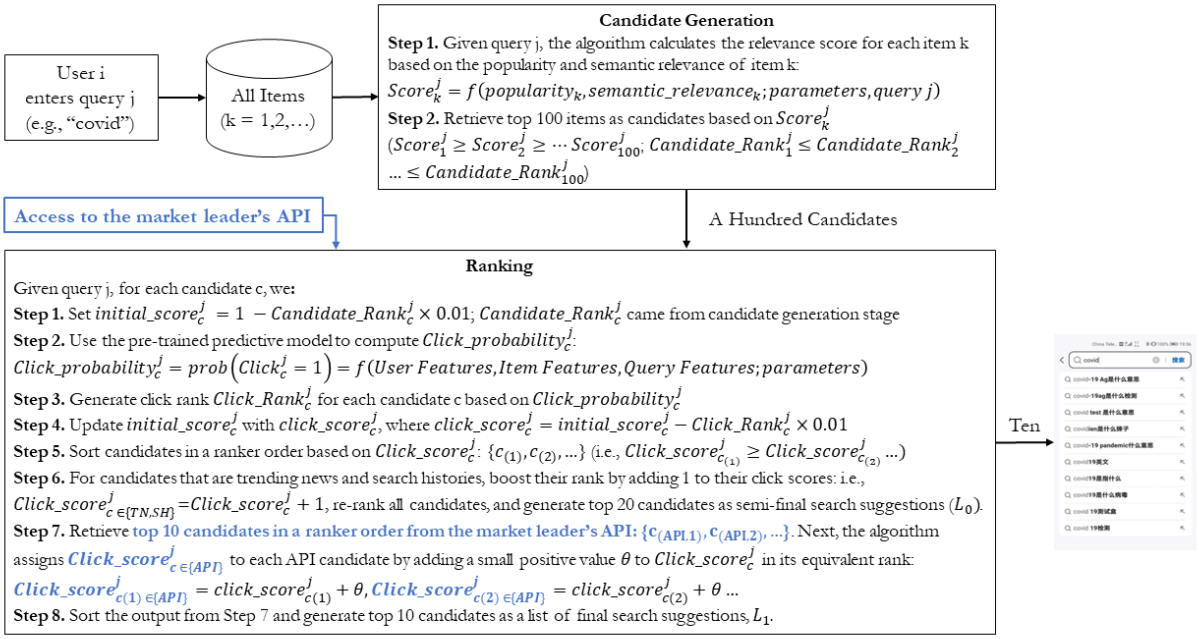
to millions of users at scale, the market leader can (1) provide the company with access to more precise candidate items with respect to their popularity and freshness, and (2) expand the scope of the company's candidate items with respect to their topic coverage, both of which can be used as high-quality inputs in the ranking stage to generate a list of search suggestions.

How does the market leader's API work? At a high level, the company and the market leader form a business-to-business contractual relationship where the company acquires an exclusive license to the market leader's API. To initiate the data request, the company provides the market leader's API with its user-submitted queries in real time, which are used by the market leader's API to return a set of candidate items. Importantly, these candidate items are **depersonalized** by design because the company provides the market leader's API with only query terms rather than users' personal data. And then, at the ranking stage, the algorithm uses (1) candidates items from the market leader's API as an additional input along with (2) its own candidate items to generate a list of search suggestions in a ranked order. Notably, we believe this setup reflects how data-sharing agreements between companies would look in practice; it provides an example of how gatekeepers could provide smaller players with "access on fair, reasonable and non-discriminatory terms to ranking, query, click and view data" if mandated within the framework of the Digital Markets Act. Lastly, the company and market leader have a revenue-sharing agreement where the company pays service fees to the market leader at a undisclosed rate, the market leader compensates the company through the volume of search query terms (see Figure A1).

Concretely, building on Figure 2 and the discussion above, Figure 3 decomposes the entire process where the company's algorithm meets the market leader's API. To fix ideas, consider the user i who enters the search query j into the search box (e.g, covid). First, the company's candidate generation algorithm retrieves about 100 candidate items (indexed by c) in a ranked order ($Candidate_Rank_c^j$) from its item base (items indexed by k=1,2,3,...,K) based on their relevance to query j and their popularity (i.e., Steps 1 and 2 under **Candidate Generation**). Second, the company's ranking algorithm scores each candidate item based on user (e.g., location, search histories), query (e.g., topic, freshness), and item features and sort them according to their predicted click score $Click_Score_{c,API}^j$ (i.e., Steps 1 through 6 under **Ranking**). Third, given query j, another

set of depersonalized candidate items is retrieved from the market leader's API in a ranked order ($c_{API(1)}$, $c_{API(2)}$...). Fourth, for each candidate item from the market leader's API, the ranking algorithm assigns a score, $Click_Score^{j}_{c,API}$, which adds a small positive value θ to the click score of the company's own candidate item in the equivalent rank to boost the rank of the marker leader's candidate items (i.e., Step 7 under **Ranking**). Lastly, the ranking algorithm re-ranks the mix of candidate items from two sources to generate a list of highest-scoring items as search suggestions (i.e., Step 8 under **Ranking**). In summary, access to the market leader's API changes the supply of candidate items entering into the ranking stage. Whereas the market leader's candidate items receive a relative rank priority, they are scored after the company's own candidate items are scored.

Figure 3: Search-Suggestion Generating Process: Decomposing the Algorithmic Funnel



How do we evaluate the performance of search suggestions? Our conversations with the search product team revealed that the target metric is click-through rate (CTR), a widely-used key performance indicator of search success in online markets. This is particularly true for this search product since the aim was to monetize these suggestions through sponsored links. As an example,

the team experimented with 'direct reach cards' that would show up in search suggestions and would lead the user directly to a website in line with the predicted interest. In the context of search suggestions, CTR is measured as the ratio of number of clicks to search suggestions in a list to the number of exposures. Specifically, when a user starts typing a keyword into the search bar, this user is exposed to a list of search suggestions in this session (i.e., one exposure). If a user clicks to any search suggestion in this list, such an action will be counted as one click. To ensure that our results are not sensitive to the variation in exposures, we conduct robustness checks using alternative measures such as the number of clicks and the probability of any click. We can measure CTR for each user on a daily basis or over a longer period (e.g., week). To capture the overall user activities rather than a user's specific search session, we focus on the *aggregate* CTR for each user: the ratio of the total number of clicks to total number of exposures over the same period of time.

Yet measuring the value of the market leader's data (candidate items) on target metrics such as CTR is challenging due to the lack of exogenous variation. To circumvent this challenge, we conducted a large-scale field experiment where we manipulated access to the market leader's API while holding all other aspects (e.g., algorithms, user interface) equal. Therefore, we can cleanly identify the impact of the market leader's data on search-suggestion CTR.

3 Field Experiment

3.1 Experimental Design

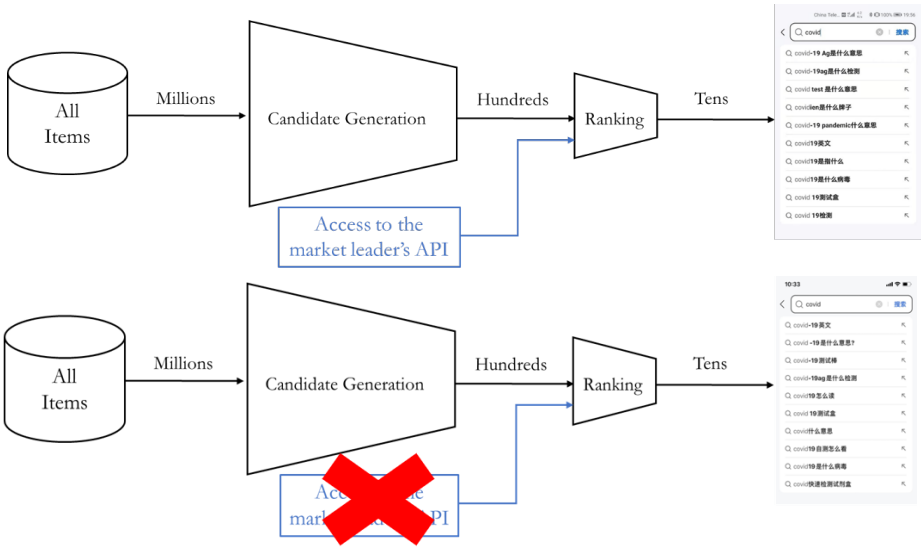
The experiment was conducted on the company's centralized experimentation platform, an overlapping experiment infrastructure designed to run experiments simultaneously and efficiently. Following the standard design of overlapping experiment (e.g., Figure 2b in Tang et al. (2010)), the company ran this longer-term experiment in a non-overlapping domain where users do not overlap with users in the experiments running simultaneously in domains for any other in-app products such as news feed and eBooks streaming. The randomization was implemented at the user level: when they started typing keywords into the search bar in a given time point during the experiment, these users were randomly assigned to one of two conditions. Once a user was assigned to a condition,

this user stayed in the same condition until the experiment ended. We maintained a consistent treatment assignment for 108 days (about 16 weeks) from May 2021 to September 2021.

Control (N=1,194,619): In response to user-submitted queries, search suggestions are generated by the company's proprietary algorithm funnel, including item generation, candidate generation, and ranking. Concretely, at the ranking stage, the ranking algorithm scores candidate items from two data sources: including (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. Next, the ranking algorithm generates the highest scoring items in a ranked list. As a result, users in this condition see a sizable proportion of search suggestions supplied by the market leader even though the number of search suggestions from the market leader might differ across users depending on queries, demographics, and search histories.

Treatment (N=1,195,625): In response to user-submitted queries, search suggestions are generated by the same proprietary algorithm funnel, which does not rely on the market leader's API at the ranking stage. The ranking algorithm (the same as the one used in the control condition) scores only candidate items retrieved from the candidate generation stage and generates the highest scoring items in a ranked list (i.e., there is no Step 7 in Figure 3). As a result, users in this condition never see search suggestions supplied by the market leader. Figure 4 visualizes our design.

Figure 4: Experimental Design



3.2 Data

We had access to data at the user-day level where we observe a unique user identifier, treatment status, number of exposures to search suggestions, number of clicks to search suggestions, login status, and search button usage. In addition, we observe pre-experimental characteristics such as demographics (e.g., gender, city of residence), mobile operating system (e.g., Android), and activity level (e.g., active days in the past 30 days). The main analysis relies on data aggregated from user-day level to user-level across the length of the experiment (i.e., the unit of observation in our data is at the user level). The dependent variable, aggregate CTR, is computed for each user as the ratio of the total number of clicks over the entire experiment to total number of exposures over the entire experiment. We conduct several checks to ensure that the random assignment is successful. Table 1 shows the mean difference in each variable across conditions is neither economically nor statistically significant. Figures A2 and A3 show that users were equally likely to be assigned to either condition regardless of when they first logged into the app and used the search bar.

Table 1: Randomization Checks

User Characteristics	Control	Treatment	p-value
Male	0.503	0.504	0.672
Larger Cities	0.505	0.505	0.425
Smaller Cities	0.434	0.435	0.349
Mobile Operating System: Apple iOS	0.107	0.107	0.736
Mobile Operating System: Android	0.837	0.837	0.794
New User (overall in-app activities; 1=yes)	0.110	0.110	0.666
Active days in the past 30 days (search activities)	100	99.826	0.426
Query views in the past 30 days (search activities)	100	99.383	0.126

Notes. Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, values reported in the last two rows were normalized so that the variable means in the control condition are 100.

4 Empirical Analyses

4.1 Empirical Framework

We use potential outcome framework to describe our model. For illustration purposes, we consider an experiment with N users who are randomly assigned to one of two conditions: i.e., treatment (e.g., API removal) or control condition. For a set of independent and identically distributed users $i = 1,...,n$, we observe the outcome of interest Y_i (e.g., CTR aggregated over 108 days), treatment assignment T_i, and a vector of user characteristics Z_i (e.g., demographics, past search activities). For each user i, there are two potential outcomes: if a user is assigned to the treatment condition, we observe the outcome $Y_i = Y_{i1}$, and if the user is assigned to the control condition, we observe $Y_i = Y_{i0}$. In theory, the average treatment effect (ATE) is $E[Y_{i1} - Y_{i0}]$, which can be used to assess whether the treatment causes changes in Y_i. Alternatively, we can estimate the following regression:

$$Y_i = \alpha + \beta \times T_i + \epsilon_i, \tag{1}$$

where β captures the causal effect of the removal of the market leader's API on the outcome of interest (e.g., CTR aggregated over 108 days). Because of the successful randomization, we do not expect the controls Z_i to affect the estimate of β. Therefore, we estimate the regressions without control variables as the baseline results and use heteroskedasticity-robust standard errors.

Another important consideration is that the regression coefficient β itself does not provide a sense of the effect magnitude. Therefore, we report the estimates as *lift* to facilitate the interpretation of estimates as the magnitude and comparison across experiments (Gordon et al., 2022): the incremental CTR among treated users relative to control users as a percentage of CTR among control users. A positive (negative) value of the estimated lift indicates an increase (a decrease) in CTR among treated users relative to control users. Importantly, the lift is a ratio of two random variables, Y_1 and Y_0, making it a random variable. We use the Delta method to derive approximations for the mean and variance of the lift:

$$\mathrm{E}(\text{lift}) = \mathrm{E}(\frac{Y_1 - Y_0}{Y_0}) = \mathrm{E}(\frac{Y_1}{Y_0}) \approx \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\mu}_0}, \tag{2}$$

$$\text{Var(lift)} = \text{Var}(\frac{Y_1 - Y_0}{Y_0}) = \text{Var}(\frac{Y_1}{Y_0}) \approx (\frac{\hat{\mu}_1}{\hat{\mu}_0})^2[(\frac{\hat{\sigma}_{\hat{\mu}_1}}{\hat{\mu}_1})^2 + (\frac{\hat{\sigma}_{\hat{\mu}_0}}{\hat{\mu}_0})^2], \quad (3)$$

where Y_1 and Y_0 are independent random variables because no user is in both treatment and control conditions. $\hat{\mu}_1$, $\hat{\mu}_0$, $\hat{\sigma}_{\hat{\mu}_1}$, and $\hat{\sigma}_{\hat{\mu}_0}$ refer to the treatment sample mean, control sample mean, standard error for the treatment sample mean, and standard error for the control sample mean, respectively (e.g., , $\hat{\mu}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{i1}$).

4.2 Baseline Results

Our starting point is estimating the average treatment effect. There is a significant and negative average treatment effect: API removal leads to a decrease in CTR by 4.6% (±0.3%) in the treatment condition relative to the control condition (i.e., the average lift is –4.6% with the 95% confidence interval in parentheses). The large sample size allows us to reach considerable precision such that the width of the 95% confidence interval is about 10% of the absolute value of the point estimate. The magnitude is economically significant especially in the search context.[5] We benchmark the magnitude in a couple of ways in Section 4.5 and find it to be economically meaningful and in line with some other studies in the broader literature.

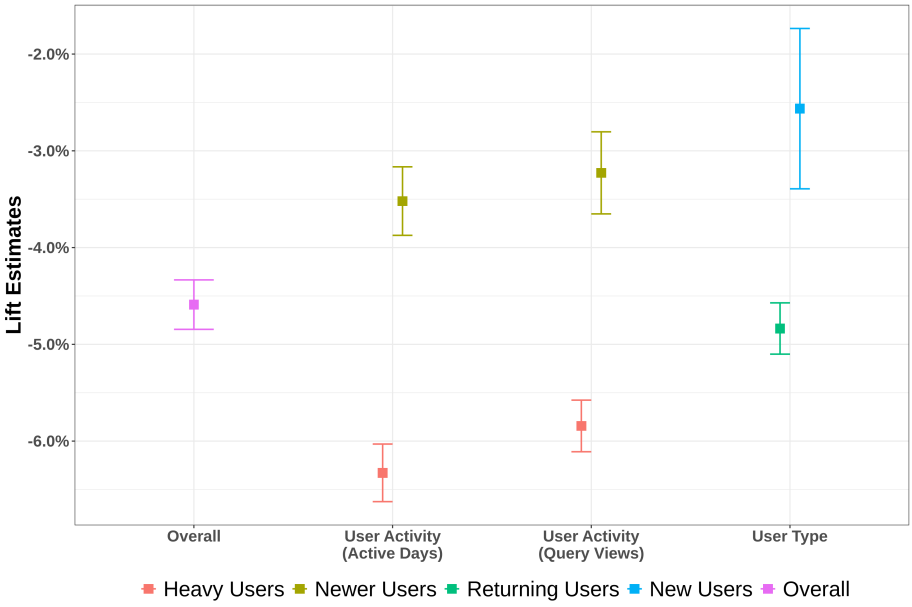
To dig into *who* is more (less) responsive to the removal of external data, we explore heterogeneous treatment effects along user characteristics. Specifically, we examine whether users with different activity levels are more or less responsive to the market leader's data as additional candidate items into the ranking algorithm. A priori, it is unclear whether the additional data will help solve the cold start problem for newer users (i.e., customer acquisition) or provide a more streamlined experience for heavy, more active users (i.e., customer retention), or both. To examine this, we use the number of search-related active days in the past month prior to the experiment to differentiate newer users from heavy users. We construct an indicator variable, heavy user, which is equal to 1 when a user's number of search-relate active days is strictly above the median and zero otherwise. Figure 5 shows that the magnitude (absolute value) of the negative effect of API removal

[5]See more details here Kohavi et al. (2013) and comments by Ronny Kohavi, former VP of Analytics and Experimentation at Bing, here https://www.facebook.com/watch/?v=2368597899925946.

among heavy users is 6.3% (±0.3%). Such magnitude is significantly larger than that among newer users, 3.5% (±0.4%). This pattern is highly consistent when we use alternative measures such as the number of query views in the past month prior to the experiment (1 if strictly above the median and zero otherwise) and an indicator of whether a user is a returning customer based on pre-experimental data.

These results highlight that relative to infrequent or newer users, users with heavier search activities in the pre-experimental period are more responsive to the removal of the market leader's data. A plausible explanation is that because heavy users generate more data on the platform and experience quality recommendations more often, they may set a higher expectation of product quality and are more sensitive to changes in the quality of search suggestions due to the absence of the market leader's data. In this regard, our results complement Sun et al. (2021) where the authors find that the use of personal data in the recommendation benefits newer users more when data volume and customer resilience coexist in their context.

Figure 5: Average Treatment Effect and Heterogeneous Treatment Effects by User Characteristics



Notes: Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users (e.g., the average treatment effect is –4.6%). Error bars represent 95% confidence intervals.

4.3 Treatment Effects by Query Type

To dig into *how* and *why* external data affects user engagement, we explore how treatment effects vary by the type of query. Ideally, we would like to observe the logs of all users in the experiment such that we have access to every query term entered by a user, corresponding content category associated with the query, a list of search suggestions exposed to a user, and whether a user clicks, and what a user clicks. Yet we were only able to collect the logs of users that clicked at least once where we observe the query terms entered by a user, corresponding content categories associate with such terms, and what a user clicked. In the remainder of this section, we used the logs of this subsample of users to construct our variables of interest.[6]

First, we examine whether the market leader's data affects what users search. To do so, we use the total number of unique content categories across the entire experiment as the dependent variable. Column (1) of Table 2 shows that API removal reduces the number of content categories searched, suggesting that the market leader's data enhances the company's product performance by expanding the breadth of topics that the users can search for.

Table 2: Treatment Effects by Query Type

	(1)	(2)	(3)	(4)	(5)
Variables	Unique Cat.	CTR Explicit	CTR Implicit	CTR Popular	CTR Niche
API Removal	-0.0297***	-0.0071***	-0.1085***	-0.0164***	-0.0065
	(0.0014)	(0.0015)	(0.0036)	(0.0016)	(0.0037)
Observations	1,631,260	1,631,260	1,631,260	1,415,958	1,415,958

* $p<0.05$, ** $p<0.01$, *** $p<0.001$. Standard errors in parentheses. This analysis was conducted with the subsample of users that clicked at least once over the course of the experiment.

Second, we examine whether the effect of API removal varies by the clarity of users' search intent. When users enter a query term with an explicit search intent (e.g., Covid), the company's natural language processing tool is able to classify the topics of such terms into 31 first-level categories (e.g., Health) and 167 second-level categories (e.g., Health-Disease). When users enter query terms

[6]We verified that the user demographics and usage activities are balanced among users in the treatment and control conditions in this subsample (for details, see Table A1).

with implicit search intent, the company's natural language processing tool classifies such terms as null or other. Therefore, we construct an indicator variable, explicit search intent, which is equal to 1 if the content category of a user's query can be explicitly labeled by the tool (e.g., Covid: Health-Disease) and 0 if it is classified as null or other. To understand how the NLP tool categorizes explicit vs. implicit search intent, it is important to understand the how queries in Chinese are usually typed using an input method called Pinyin. Pinyin is a collection of input codes for Chinese, which is similar to the Phonetic symbols for English. A user can either enter a Chinese character or a Chinese word in one round into the search box by typing Pinyin in the keyboard. When users' search intent is explicit, they will likely type the input codes and enter the full length queries and can be classified into a category by NLP tools. When users' search intent is implicit, they will type the Chinese character one by one also resulting in slow typing speed making it hard for NLP tools to classify them. We compute the CTR among query terms with explicit search intent and CTR among those with implicit search intent for each user each day, aggregate them across the entire experiment at the user level, and use them as two separate dependent variables. As before, Table 2 reports the treatment effect estimates as lift. Table 2, Columns (2) and (3) show that the decrease in CTR due to API removal is much higher when users' search intent is implicit. In other words, the market leader's data is more valuable when users' search intent is implicit. A plausible explanation is that search suggestions based on real searches by all users on the market leader's platform are more likely to capture users' interests when they are less explicit about their queries.

Third, we examine whether the effect of API removal varies by query popularity conditional on explicit search intent. We conduct this analysis with search records of users that clicked at least once and those with explicit content categories (i.e., content categories are neither null nor other). Following the literature around the Pareto principle (Brynjolfsson et al., 2011), we define the content categories that generate 80% of the clicks in the control condition as popular (mainstream) content (e.g., Education-K-12, Books-Novel, Health-Disease) and those that does the rest of 20% as niche content (e.g., Education-Study Abroad, Music-Music Radio, Government Affairs-Nonprofit Organization). Columns (4) and (5) suggest that the reduction in CTR is driven by relatively popular (mainstream) topics. This suggests that the proprietary data from the market leader helps

identify certain 'easy hits' that are popular with the control group. A plausible explanation is that because the queries sent to the market leader's API are depersonalized, candidate suggestions generated by the market leader's API capture popular content based on real searches by all users on the market leader's platform that are relevant to user-submitted queries. In contrast, clicking niche content is not affected by API removal because recommending niche content relies more on personal data. This demonstrates that the impact of API removal on product performance is driven by mainstream content that requires less personalization. This has implications for managers and policy makers that are looking to (privacy-preserving) data sharing regulations across companies.

4.4 Long(er)-Term Effects

A novel aspect of our design is that a consistent 16-week treatment assignment allows us to estimate longer-term effects. Following Huang et al. (2018), we estimate a regression for each week of the experiment as if that week were the final one. This gives us 16 separate regression estimates. Figure 6 shows how lift estimates vary over time, with the solid line representing the point estimates and the dotted lines representing the 95% confidence intervals. The estimates start at about a 8.1%–9% decline, decreasing to 3.6%–4.5% in the last few weeks. Thus, had we run a short-term experiment, we could have overestimated the value of the market leader's data by a factor of 2.

With such a longer-term analysis, we need to ensure that the estimates are not driven by differential attrition across treatment and control conditions. If there were users becoming inactive in terms of the overall product usage, the estimates would capture the treatment effect due to temporal changes in the sample of users. To alleviate this concern, we examine whether the treatment induces a significant change in users' login activities. On average, API removal does not cause changes in the total number of days a users logged in and used the search product (lift = –0.0028, n.s.). We plot the weekly estimates with respect to the impact of API removal on the number of login days in a week. Figure 7 shows that API removal does not cause a significant change in login activities among treated users relative to control users in any week (confidence intervals consistently contain zero). Together, these results suggest that the decrease in the magnitude of lift estimates is unlikely to be driven by differential attrition across two conditions.

Figure 6: Longer-Term Treatment Effects

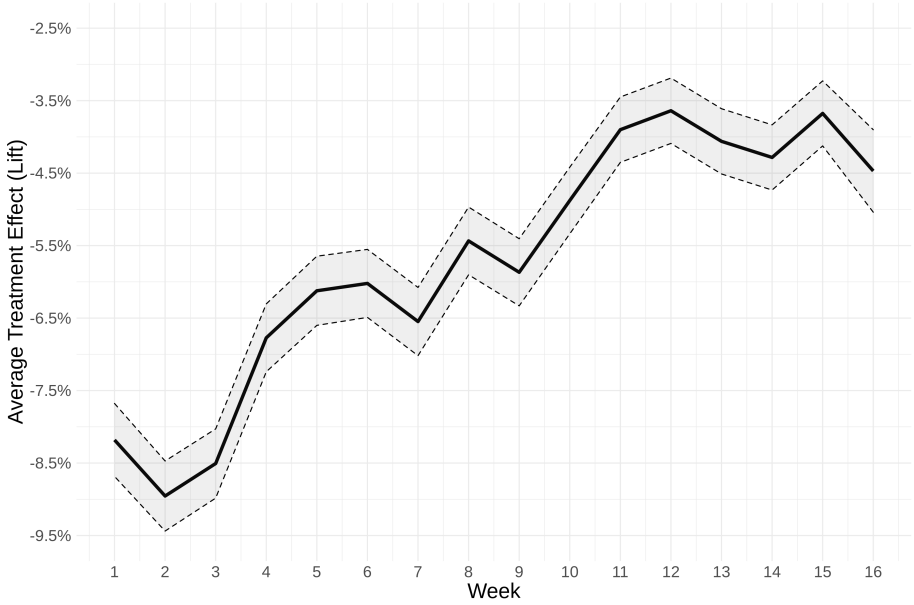
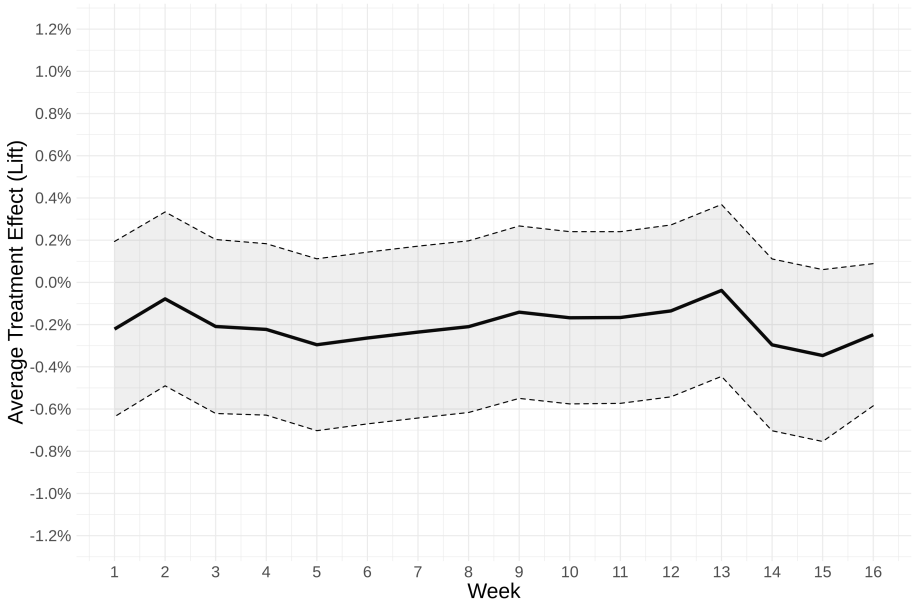


Figure 7: Longer-Term Effects of API Removal on the Number of Login Days



Lastly, we look at another target metric, search button usage, to explore whether changes in CTR on search suggestions can be compensated by the utilization of an alternative search function. On average, API removal significantly increases the usage of search button by 5.7% ($\pm$1.1%). Figure A4 in the Appendix shows how the estimated lift in search button usage vary over time, and they suggest that the demand spillover to search button remains positive over the 16-week period.

Collectively, there are three takeaways from this analysis. First, a short-term evaluation would have made us overstate the value of the market leader's data. Second, we provide suggestive evidence that users can adapt their behavior as the market leader's data is removed over a prolonged period of time. Third, the positive impact of API removal on search button usage suggests a positive demand spillover to search button as a substitute for search suggestions.

4.5 Economic Implications

To understand the economic implications of access (or lack thereof) to external data, we carry out a back-of-the-envelope computation. We focus on translating the decrease in CTR into revenue loss suffered by our partner company. To carry out this exercise, we leveraged internal data to put a monetary value on the decline in CTR. Specifically, the unit query revenue is 0.012 Chinese yuan (CNY). The unit query revenue estimate is obtained from an internal one-month experiment intended to measure the return on investment of a search-activity promotion policy. The unit query revenue is on the lower end of related industry estimates exactly because the search product is a new one with limited monetization. Hence, our revenue calculations should be thought of as a lower bound. The experiment lasted for one month so we aggregate the revenue to query-month level. An average user performs 7.25 search-suggestion clicks per month. For the 1,195,122 users in the treatment group in the experiment, a 4.6% decrease in search-suggestion CTR creates a loss of 4,782 CNY per month (i.e., $7.25 \times 0.012 \times 4.6\% \times 1,195,122$). After we extrapolate the estimates to monthly active users of search products (about 100 million), the loss is over 400,200 CNY per month or 4.8 million CNY per year. This extrapolation to 100 million monthly active users is not different from other global contexts. For example, in March 2019, Bing had 126 million active users even though it accounted for about 2% of the search market globally. Therefore. this exercise suggests

that there is a bang for the buck for the company by leveraging market leader's proprietary data.

In addition to these computations, the magnitude of the treatment effect estimate in our experiment can be benchmarked to other studies in the literature. The average effect of 4.6% is comparable to Berman and Israeli (2022) showing that descriptive analytics increases revenue between 4%-10% and Brynjolfsson and McElheran (2016) showing that data driven decision making increases firm performance by about 3%. Bar-Gill et al. (2021) shows that providing access to a data dashboard lead to an increase in sales by 3%. Finally, Kim (2019) demonstrates that providing access to competitor information leads companies to increase sales by 8%. Our average estimate as well as its evolution across the four months of the experiment falls within the range of the estimates mentioned above. Indeed, the studies analyze the impact of additional data or information provided to a firm or seller which is in line with the framework of the current study.

Overall, this section allows us to pin down the economic significance of the magnitudes which can have implications for the focal company but also policy makers.

4.6 Robustness Checks

We carry out a variety of checks to test for the robustness of our results (see Table A2 in the Appendix). First, the treatment status on the first day could influence whether users come back more often or not. Therefore, we estimate the treatment effect using only the first-day observation of each user in the experiment (i.e., the first day of the experiment when a user interacts with search suggestions) so that we avoid potential feedback loops and path dependency in user behavior. Second, we estimate a linear regression adjusting for observed user characteristics to account for imbalance in such characteristics across treatment and control conditions (if there is), potentially improve the precision of the estimate, and check the stability of results among the sample where all user characteristics are observed. Columns 1 and 2 show that the results are qualitatively similar to our baseline estimates with a negative and statistically significant effect. Third, we use alternative operationalizations of the dependent variable, including the click dummy (Column 3) and the logarithm of one plus clicks (Column 4). Our results are robust and qualitatively similar. Lastly, we estimate a model with all moderators in the same linear regression rather than subsample

analysis reported in Figure 5. The results are very consistent (Column 5). In summary, these checks provide an additional degree of confidence in our baseline results.

5 An Extension with Field Experiment 2

In the previous section, we have provided evidence on the impact of the market leader's data on the company's search product performance. In this section, we leverage another field experiment in July 2021 that ran for one day[7] to achieve three objectives: (1) replicate the main results from the first field experiment, (2) validate the manipulation of search suggestions provided by the market leader's API, and (3) assess the relative impact of manipulating the supply of candidates into the ranking algorithm versus manipulating the rank directly in the same experimental setup.

5.1 Experimental Design

Similar to field experiment 1, the randomization in this experiment was implemented at the user level (see the randomization checks in the Appendix, Table A3). A total of 250,281 users were randomly assigned to one of three conditions:

Control (N=83,500): This condition is similar to the control condition in field experiment 1. Search suggestions are generated by the proprietary algorithm funnel. At the ranking stage, the ranking algorithm scores candidate items from two data sources: (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. The ranking algorithm scores such candidate items and generates highest-scoring search suggestions in a ranked order.

Rank Adjustment (N=83,517): Identical to the control condition, at the ranking stage, the ranking algorithm scores candidate items from two data sources: (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. In contrast to the control condition, the ranking algorithm lowered (boosted) the rank of the market leader's candidate items (the company's own candidate items) and then generated the final search suggestions. Recall Step

[7]Although the date of this experiment overlaps with the field experiment 1, the interference between two experiments is not a concern because the company conducted the field experiment 1 in a non-overlapping domain where users do not overlap with the users in this experiment.

7 in Figure 3: the manipulation was achieved in a way that the ranking algorithm added a fixed positive value (larger than θ) to the click scores of the company's own candidate items. As a result, relative to those in the control condition, users in this condition are less likely to see search suggestions supplied by the market leader in the final list.

Removal of Access to the Market Leader's API (N=83,264): This condition is similar to the treatment condition in field experiment 1. The ranking algorithm (the same as the one used in the control condition) scores only candidate items retrieved from the candidate generation stage and generates highest-scoring search suggestions in a ranked order. As a result, users in this condition never see search suggestions supplied by the market leader.

5.2 Results

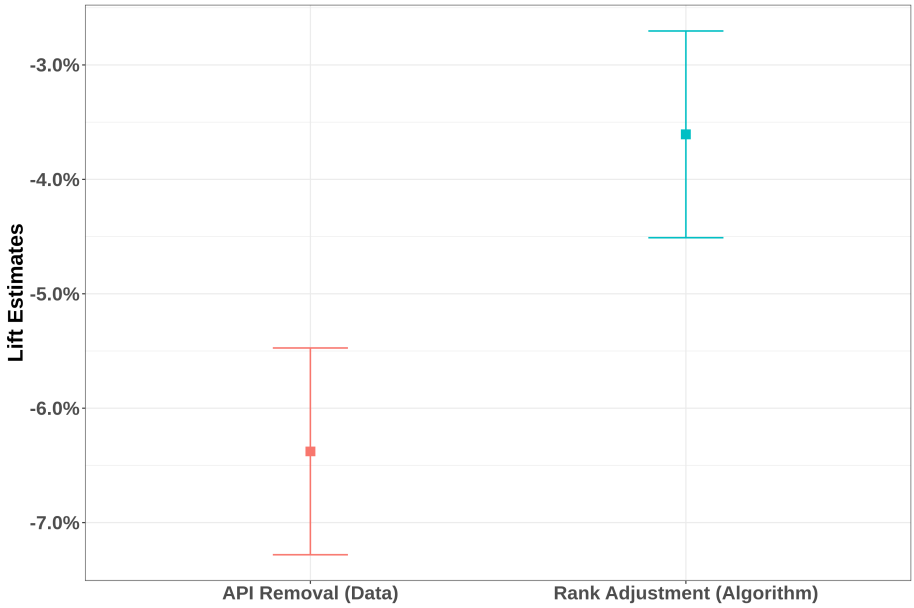
Figure 8 shows that the removal of the market leader's API leads to a decrease in CTR by 6.4% (±0.9%). This estimate is between the aggregate estimate (i.e., 4.6%) and first-week estimate (i.e., 8.1%) from field experiment 1, replicating the main effect reported earlier.

Turning to rank adjustment, we find that pushing down the ranking of search suggestions from the market leader leads to a decrease in CTR by 3.4% (±0.9%). This estimate supports that manipulating the rank directly has a significant impact on product quality. In practice, such an adjustment broadly relates to Google's adjustment in its ranking algorithms. Specifically, Google's white paper notes that "where our algorithms detect that a user's query relates to a "Your Money or Your Life (YMYL)" pages topic, we will give more weight in our ranking systems to factors like our understanding of the authoritativeness, expertise, or trustworthiness of the pages we present in response (Google, 2019)."

Taken together, these estimates provide suggestive evidence that in our context, manipulating the rank of search suggestions induces a smaller impact on CTR relative to the removal of the market leader's data as candidates for the ranking algorithm. Another takeaway is that these estimates shed light on the strength of the manipulation across different conditions. Conceptually, the manipulation of pushing down the rank of search suggestions from the market leader should be weaker than completely removing the market leader's candidate items. Therefore, the magnitude

of the estimates (6.4% versus 3.4%) increases our confidence in the success of the manipulation.

Figure 8: The Relative Impact of API Removal and Rank Adjustment



Notes: N= 250,281. Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users. Error bars represent 95% confidence intervals.

6 Conclusion

We leverage a large-scale field experiment where we randomize access to the market leader's data as an algorithmic input to generate search suggestions for users. We find that without access to the market leader's data, users click on search suggestions 4.6% less relative to those users in the condition where the algorithm has access to the external data through the market leader's API. The lack of external data leads to a reduction in the breadth of content consumed as measured by the number of content categories. These content categories are mainstream (rather than niche) in terms of the share of clicks they get from users. The external data streamlines the experience for users that highly active in using the search suggestions tool. Using the entire length of the large-scale experiment, we document significant dynamics in how users interact with the product

when there is no access to external data. In particular, we find that users adapt and the negative effect of the API removal reduces over time. Finally, using a second (short-term) experiment, we manipulate (i) access to the market leader's API and (ii) the rank of search-suggestion candidates to ensure that it is indeed the market leaders data that leads to the effects we measure.

Our study has clear managerial implications. The results suggest that leveraging external sources of data can provide a significant return for a company especially with a new product. For search products in particular, we show that access to depersonalized search results can increase engagement. This is important because there are several search products that are launched periodically (eg. Neeva, Cliqz) that could look to this potential strategy. It could help to provide a good experience especially for existing (heavy) users of the product by increasing the breadth of (mainstream) content available to them.

We believe our results also shed light on policy issues being debated currently. The benefit derived by our partner company demonstrates that depersonalized search results being mandated by regulations like the DMA might help provide a leg up for nascent search products. The reduction in mainstream content consumption might be a result of the depersonalized suggestions also highlights the limits of catering to individual preferences in the absence of use of personal data.

Our study is not without limitations. Like other studies with field experiments, we are only able to look at one setting. It would be prudent to try and replicate the general tenor of our findings in other contexts. These could include other online products as well as use of different algorithms and data types. This would, of course, depend on whether the opportunity exist to leverage such data sharing agreements. Our study is also a partial equilibrium analysis and further research could take into account more general equilibrium dimensions. For example, there could be strategic responses from other (competing) platforms in the face of such data partnerships. Analyzing how the platform ecosystem evolves with such data partnerships or through regulations would be a fruitful next step.

References

Agrawal, A., J. Gans, and A. Goldfarb (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Bar-Gill, S., E. Brynjolfsson, and N. Hak (2021). When small businesses become data-driven: A field experiment. *Working Paper*.

Benzell, S., J. Hersh, and M. Van Alstyne (2022). How apis create growth by inverting the firm. *Working Paper*.

Beraja, M., D. Y. Yang, and N. Yuchtman (2020). Data-intensive innovation and the state: evidence from ai firms in china. Technical report, National Bureau of Economic Research.

Berman, R. and A. Israeli (2022). The value of descriptive analytics: Evidence from online retailers. *Marketing Science 41*(6), 1074–1096.

Brynjolfsson, E., Y. Hu, and D. Simester (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science 57*(8), 1373–1386.

Brynjolfsson, E. and K. McElheran (2016). Data in action: data-driven decision making in us manufacturing. *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-06, Rotman School of Management Working Paper* (2722502).

Chiou, L. and C. Tucker (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research.

Claussen, J., C. Peukert, and A. Sen (2019). The editor vs. the algorithm: Targeting, data and externalities in online news. *Data and Externalities in Online News (June 5, 2019)*.

Covington, P., J. Adams, and E. Sargin (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198.

Crémer, J., Y.-A. de Montjoye, and H. Schweitzer (2019). Competition policy for the digital era. *Report for the European Commission*.

Economist, T. (2017). The world's most valuable resource is no longer oil, but data.

Google (2019). How google fights disinformation. *White Paper*.

Gordon, B. R., R. Moakler, and F. Zettelmeyer (2022). Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*.

Huang, J., D. H. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio. *IO: Empirical Studies of Firms & Markets eJournal*.

Kim, H. (2019). The value of competitor information: Evidence from a field experiment. Technical report, Working Paper Harvard Business School.

Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1168–1176.

Nagaraj, A. (2022). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science 68*(1), 564–582.

Nandy, P., D. Venugopalan, C. Lo, and S. Chatterjee (2021). A/b testing for recommender systems in a two-sided marketplace. *Advances in Neural Information Processing Systems 34*, 6466–6477.

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

Rubin, E. and A. Rubin (2021). On the economic effects of the text completion interface: empirical analysis of financial markets. *Electronic Markets 31*, 717–735.

Sun, T., Z. Yuan, C. Li, K. Zhang, and J. Xu (2021). The value of personal data in internet commerce: A high-stake field experiment on data regulation policy. *Available at SSRN 3962157*.

Tang, D., A. Agarwal, D. O'Brien, and M. Meyer (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 17–26.

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.

Valavi, E., J. Hestness, N. Ardalani, and M. Iansiti (2020). Time and the value of data. *Working Paper*.

Xue, L., P. Song, A. Rai, C. G. Zhang, and X. Zhao (2019). Implications of application programming interfaces for third-party new app development and copycatting. *Production and Operations Management*, 1887–1902.

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science 66*(3), 1045–1070.

Appendix A

Figure A1: How Does the Market Leader's API Work?

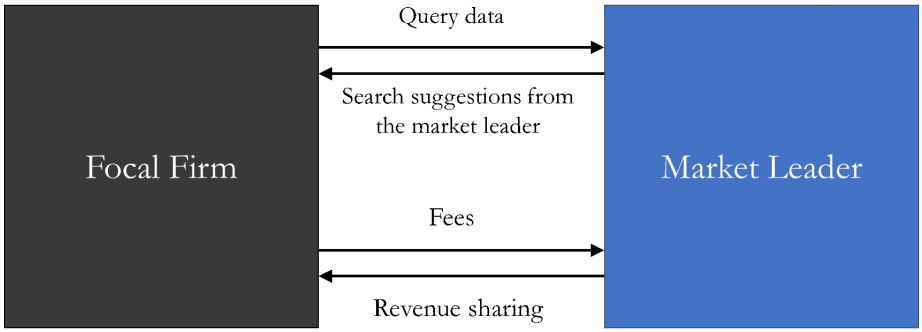


Figure A2: Proportion of Users Assigned to the Treatment Condition Over Time

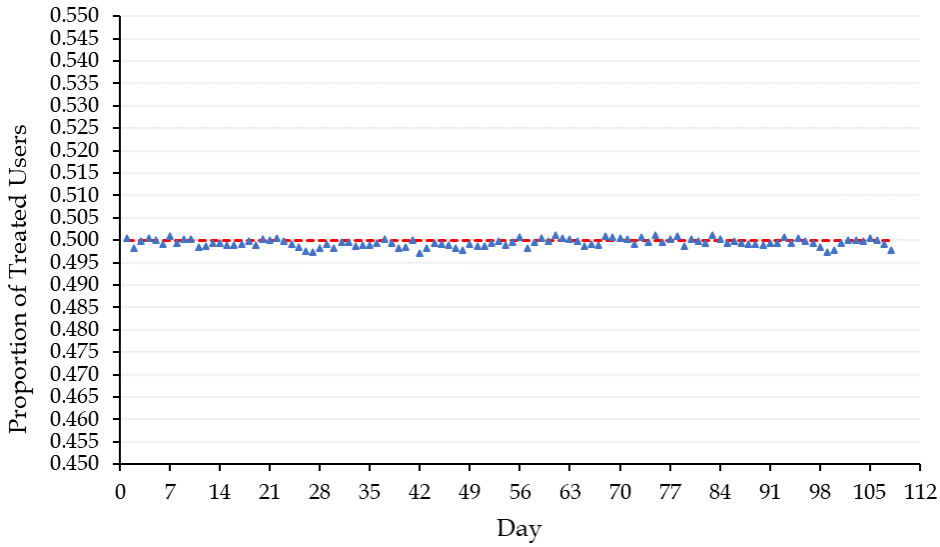


Figure A3: Proportion of New Users Assigned to the Treatment Condition Over Time

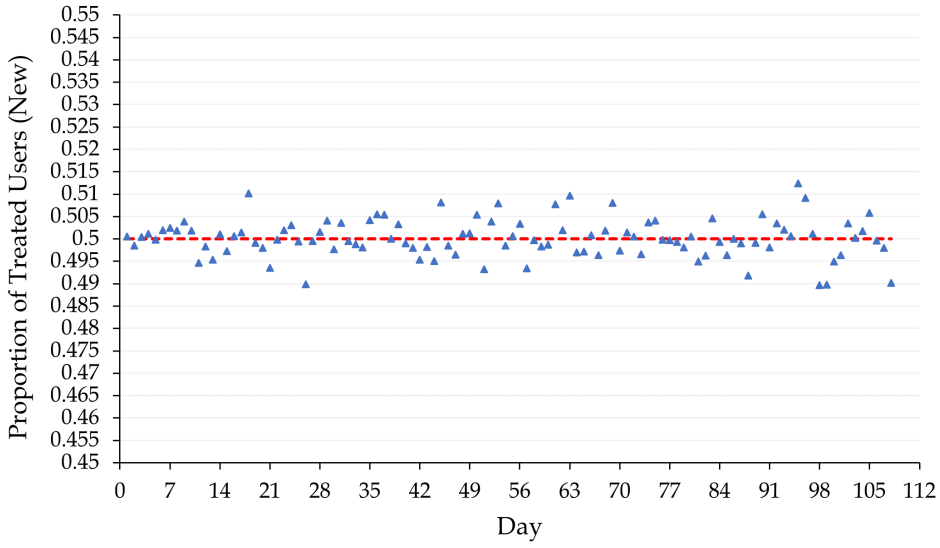
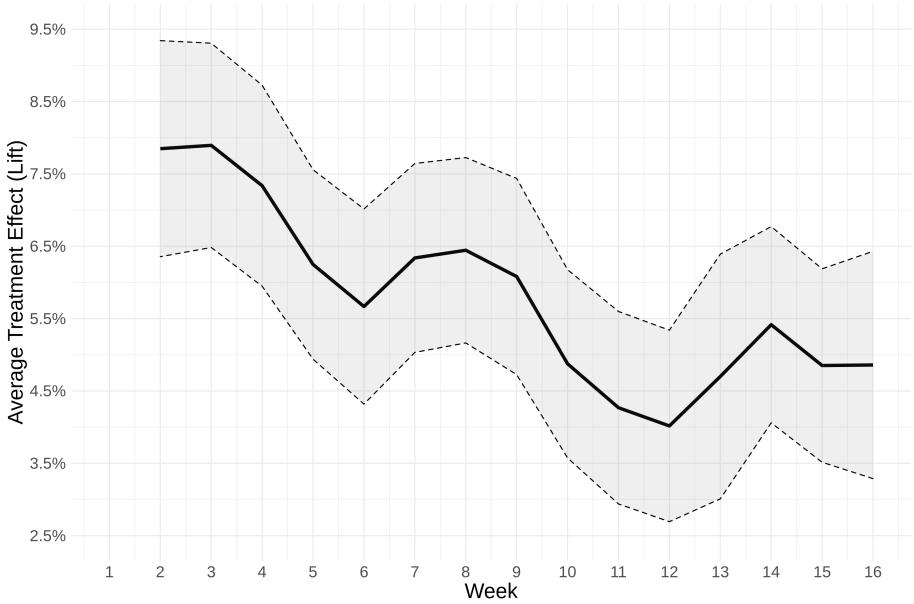


Figure A4: Search Button Usage



Notes: Lift refers to the incremental search button usage among treated users relative to control users as a percentage of search button usage among control users. Error bars represent 95% confidence intervals. Due to instrument errors, we did not have accurate log records to construct the search button usage in the first week.

Table A1: Randomization Checks (Subsample of Popular and Niche Content)

User Characteristics	Control	Treatment	p-value
Male	0.530	0.530	0.300
Larger Cities	0.553	0.552	0.209
Smaller Cities	0.438	0.439	0.156
Mobile Operating System: Apple iOS	0.125	0.125	0.503
Mobile Operating System: Android	0.869	0.869	0.336
New User (overall in-app activities; 1=yes)	0.102	0.102	0.498
Active days in the past 30 days (search activities)	100	99.905	0.688
Query views in the past 30 days (search activities)	100	99.525	0.263

Notes. Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, values reported in the last two rows were normalized so that the variable means in the control condition are 100.

Table A2: Robustness Checks

Variables	(1) Lift (First Day)	(2) Lift (Controls)	(3) Click Dummy	(4) Log(1+Clicks)	(5) CTR
API Removal	-0.0459***	-0.0539***	-0.0103***	-0.0329***	-0.0124***
	(0.0014)	(0.0013)	(0.0009)	(0.002)	(0.001)
API Removal×Female					-0.0016
					(0.001)
API Removal×Smaller Cities					0.0013
					(0.001)
API Removal×New User					0.0068**
					(0.002)
API Removal×Active Days					-0.0085***
					(0.001)
API Removal×Query Views					-0.0104***
					(0.001)
Observations	2,388,377	1,932,886	2,390,244	2,390,244	1,932,886

* $p<0.05$, ** $p<0.01$, *** $p<0.001$. Standard errors in parentheses.

Table A3: Randomization Checks: Field Experiment 2

User Characteristics	Control (C)	Algorithm (T1)	Data (T2)	p-value Diff (T1,C)	p-value Diff (T2,C)
Male	0.571	0.567	0.567	0.073	0.076
Larger Cities	0.633	0.631	0.633	0.523	0.771
Smaller Cities	0.329	0.328	0.327	0.615	0.298
New User (overall in-app activities; 1=yes)	0.010	0.009	0.009	0.699	0.496
Active days in the past 30 days (search activities)	100	100.053	99.866	0.861	0.657
Query views in the past 30 days (search activities)	100	100.436	100.399	0.547	0.577

Notes: Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, numbers in the last two rows were normalized so that the variable means in the control condition are 100.

Appendix B Theoretical Framework (In Progress)

B.1 Model

Building on Berman and Katona (2013) and Bergemann et al. (2022), we outline below a simple model to guide our empirical analyses. We assume there is a monopolistic search engine that provides search suggestions to users by displaying a limited number of phrases and/or sentences related to the queries entered by users. These phrases and/or sentences are originally from one of two sources: (1) internal data and algorithm and (2) external data, i.e., the market leader's API. Whenever a user enters the search query, the search-suggestion system ranks the related phrases and/or sentences according to a scoring mechanism and presents one from internal source and one from external source according to the scores. The incentives and characteristics of the search-suggestion service provider and users are described below.

Users and Data Sources: Users in our model express their search intent by entering queries into the search box. There are two types of search intent: (1) explicit search intent and (2) implicit search intent which are associated with different initial search cost.[8] To type in Chinese queries into the search box, users usually use an input method called Pinyin. Pinyin is a collection of input codes for Chinese, which is similar to the phonetic symbols for English.[9] Pinyin reflects the

[8]The term *initial search cost* means users' original search cost without the influence of search suggestions.

[9]See the Wikipedia page here https://en.wikipedia.org/wiki/Pinyin. According to MobTech research institute, 85.6% of users choose Pinyin as their usual input method in 2021. The other four input methods are voice, handwriting, stroke and Wubi. See the survey report by MobTech here http://www.199it.com/archives/1296409.html.

mapping of Chinese Phonetic Alphabet and Chinese characters and/or words. A user can either enter a Chinese character or a Chinese word in one round into the search box by typing Pinyin in the keyboard.[10] There exists a difference among Chinese and English typing method. For example, if a user want to search *sky*, she/he will type in s, k and y separately with the English typing method but is more likely to type in the input codes *skai* and get the full length word *sky* directly in the search box with the Chinese typing method. This difference offers an ideal setting to distinguish between implicit and/or explicit search intent and its corresponding high and/or low initial search cost. When users' search intent is explicit, they will type in the input codes and enter the full length queries, so the initial search cost can be low. Such full length queries are easy to be classified into one content category through the company's natural language processing techniques.[11] When users' search intent is implicit, they will type in Chinese characters one by one while thinking so the initial search cost can be high in this case. Also, because of the slow typing speed, each Chinese character they enter will bring up a list of search suggestions exposed to them. Such fragmental queries are so vague that natural language processing techniques can hardly classify it.

During the typing process, each query is associated with one exposure of a list of search suggestions. As we discussed earlier, the three-stage algorithmic funnel provides a short list of search suggestions that could match users' search intent. Following Berman and Katona (2020), the distance between users' real search intent and the company's or the market leader API's suggested phrase and/or sentence i associated with a specific query is denoted as d_i. Due to the lack of candidate items with respect to both quantity and quality, the company's search-suggestion system may not have the capability to provide the most desirable and nearest suggestions for some queries, especially those not explicit. As a result, the proprietary data from the market leader's API is introduced to reduce users' initial search cost. The decision on whether to click the list or tap the search button depends not only on the expected distribution of distances behind each phrase and/or sentence but also on the realized search cost e_i under the influence of the market leader's data. A rational user will click the search-suggestion list if she/he feels that one of those phrases

[10] A user can even type in a sentence directly in one round using advanced Pinyin tools such as double-stroke Pinyin (Shuangpin). See an introduction from Apple support service here https://support.apple.com/fr-fr/guide/chinese-input-method/cim9ecd77923/mac.

[11] To be more precise, the most representative technique is Query Intent Understanding.

and/or sentences suggested meets her/his intent.

As mentioned earlier, the market leader's API offers de-personalized candidate items. The candidate items are retrieved based on (1) their relevance to the original query and (2) their popularity based on market leader's accumulated search histories. Unlike the distance of the company's own items which is mainly extracted from semantic relevance and individual search histories, i.e., individual information, popularity is the most idiosyncratic information from the market leader's data. Popularity p_i reflects aggregate data or common knowledge from the huge crowd of market leader, i.e., aggregate information. Following Bergemann et al. (2022), we view the richness in the data structure as represented by d_i and p_i as the defining feature of internal and external data in the digital economy. In particular, the sum of d_i and p_i reflects the importance of social learning as enabled by search suggestions. Both distance and popularity are independent across phrases and/or sentences. The service fee paid to the market leader is denoted as c_i and the unit fee for each search is fixed for each API request, i.e., c_i=c.

The multiplier between the popularity p_i and the realized search cost e_i is $1/\rho$ for each phrase and/or sentence i, i.e., e_i=p_i/ρ with ρ>0. Whether the multiplier of a particular phrase and/or sentence is larger than one could be driven by several factors of the query. For example, for a query reflecting implicit intent, aggregate information offered by the market leader can predict a user's real intent and help her/him to reduce search cost, suggesting a multiplier possibly less than one. However, a multiplier larger than one is also possible because of the disturbance brought by aggregate information when intent is explicit and more precise search suggestions are needed. At the same time, the multiplier can be reduced because of the much higher number of topics covered by market leader to meet users' diverse search intent.

Search Suggestions: As a new product without much advertising, the search-suggestion service provider is on the users' side. Its goal is to provide users with phrases and/or sentences items that best match their search intent. To rank items, the search-suggestion system scores each item on its distance from search intent using individual data gathered from both internal and external sources, along with aggregate information involved in this process. Because of the lack of candidate items and efficient algorithm at the early stage, the search-suggestion system can therefore only measure

distance with an error and cannot observe it directly. We model the score of each item as

$$s_i = d_i + \alpha(p_i - c_i) + \varepsilon_i, \tag{4}$$

where α is a parameter denoting the search-suggestion service provider's API investment, which affects the final aggregate information utilized in this system. ε_i is the measurement noise, distributed according to a distribution with cumulative distribution function (c.d.f) F_ε and mean 0. If the company removes the market leader API, α will approach zero from the right as the company's collected data also contains aggregate information to some extent.

The expected effectiveness of a search-suggestion item i is then

$$U_i = d_i \cdot \Pr(s_i - s_j) + e_i^2/2, \tag{5}$$

Where the realized search cost is at a quadratic form of $e_i^2/2$. The company's goal is to minimize the objective function, i.e., $\min_{\alpha} U_i$. The utility of the user is negatively correlated with the effectiveness, so users' click-through rates and consumer welfare decrease with the effectiveness of search suggestions. To illustrate our results, we assume that the measurement error has a uniform distribution $\varepsilon_i \sim \mathrm{U}[-\sigma/2, \sigma/2]$. To show the impact of the market leader's API data on the overall ranking, we use $P(\alpha)$=$P(\alpha; \sigma, d_1, d_2, p_1, p_2, c)$[12] to denote the stability of the ranking process, which is the probability of the item with the lowest distance marked a higher score. The click-through rates are a result of the endogenous user click behaviors in equilibrium.

B.2 Value of External Data

API Effects: Let x=$\overline{s_i}$−$\overline{s_j}$, μ=d_i−d_j and p=p_i−p_j. As mentioned earlier, in the mixing stage for internal and external data, an API item is always put on a higher place than a corresponding own item. Without loss of generality, when we have two items i and j, let item i denoting the item from API and item j denoting the item from the company's own item pool. We define the case when users have implicit intent as $0<\mu<\sigma$, and the case when users have implicit intent as $\mu<-\sigma$

[12] c will be cancelled out finally because the unit fee is fixed for each API request. Then $P(\alpha)$=$P(\alpha; \sigma, d_1, d_2, p_1, p_2)$.

or $\mu>\sigma$. The logic behind such definition is when users' intent is explicit, the company's algorithm can easily identify the relative distance between items from these two sources and better utilizing individual information for personalization. However, when the intent is implicit, the algorithm is not that valid and the relative distance always falls into the ranking algorithm's measurement error range.

Proposition 1. (a) For explicit intent, (1) when the distance of API items is larger than the distance of the company's own items, a large enough $\alpha>0$ improves the stability of the ranking algorithm. This relationship only holds for any σ and $\rho=1$, or some $\sigma<1$ and any ρ, (2) however, when the distance of API items is less than the distance of the company's own items, aggregate information from the market leader's API is detrimental to the stability of the ranking algorithm, with any $\alpha>0$ that is not too large. For implicit intent, a large enough $\alpha>0$ disturbs the stability of the ranking algorithm.

(b) Suppose α is small. For explicit intent, (1) when the distance of API items is larger than the distance of the company's own items, a large enough $\alpha>0$ improves user welfare, (2) however, when the distance of API items is less than the distance of the company's own items, aggregate information from the market leader's API is detrimental to the user welfare. For implicit intent, any large enough API investment helps to increase users' click-through rates and welfare.

User Heterogeneity: To make sure the whole search-suggestion system provides an available list for each search, the API investment should be lower than a threshold. Assuming there's two types of users: heavy and newer. Compared to newer users, heavy users search more, leave more footprints and generate more data on the platform which helps the search-suggestion system to better identify users' search intent. Based on mass data of the huge crowd of market leader, heavy users' search intent is more predictable and will converge to an average while newer users' search intent is more unusual and fluctuated, so for the same piece of query, $d_{iH}<d_{iN}$.

Proposition 2. For heavy users, the threshold of API investment is higher since their needs are more predictable with API. With a large API investment in a new product, newer users sometimes can not benefit from API data due to some ranking failures and infrequent search needs. Heavy users hurt more if API is removed from such a new product.

Query Type Heterogeneity: Queries associated with explicit intent can be further divided into niche and mainstream (popular) categories. Niche queries are associated with a higher $p{=}p_i{-}p_j$. Mainstream queries are associated with a higher ρ since API data can cover a wider range of query categories.

Proposition 3. Holding p constant, under the condition where API data can improve user welfare, mainstream queries require a lower API investment than niche queries to let API data benefit users, i.e., the click-through rates of mainstream queries are more sensitive to API investments and aggregate data inputs.

Proof: Let $F_{\varepsilon_i-\varepsilon_j}$ be the cumulative distribution function (c.d.f) of a triangle distribution $\varepsilon_i-\varepsilon_j \sim \mathrm{T}[-\sigma,\sigma]$ with mean zero and $f_{\varepsilon_i-\varepsilon_j}$ being its probability density function (p.d.f). According to the details of search-suggestion generating process, since the system is built on pointwise learning to rank method[13], each phrase and/or sentence faces the following first-order condition (FOC) with respect to their scores resulting from the effectiveness function:

$$d_i{\cdot}f_{\varepsilon_i-\varepsilon_j}(\overline{s_i}{-}\overline{s_j}){=}{-}\frac{1}{\rho^2\alpha^2}(\overline{s_i}{+}\alpha c_i{-}d_i), \tag{6}$$

where $\overline{s_i}{=}E_{\varepsilon_i}[s_i]$. Let $x{=}\overline{s_i}{-}\overline{s_j}$, $\mu{=}d_i{-}d_j$ and $p{=}p_i{-}p_j$. By subtracting both FOCs and using the fact that $f_{\varepsilon_i-\varepsilon_j}$ is symmetric around zero, we can rewrite the condition as

$$f_{\varepsilon_i-\varepsilon_j}(x){=}{-}\frac{1}{\mu\rho^2\alpha^2}(x{+}\alpha c{-}\mu), \tag{7}$$

An interior solution x^* would require both FOCs and second-order conditions (SOCs) to hold as well as $-\sigma{\le}x^*{\le}\sigma$. Here, we assume $\mu{<}{-}\sigma$ or $\mu{>}0$ for the error to have any effect and the company to have an incentive to invest in API data for a new product. This condition also ensures that $-\sigma{\le}x{\le}\sigma$. When $0{<}\mu{<}\sigma$ and $\alpha^2{\ge}\frac{\sigma}{\rho^2}$, or when $\mu{<}{-}\sigma$ and $\alpha^2{<}\frac{\sigma}{\rho^2}$, the equilibrium solution is

$$s_i^*{-}s_j^*{=}x_L^*{=}\frac{-\sigma\mu\rho^2\alpha^2-\rho^2\alpha c+\sigma^2\mu}{\mu\rho^2\alpha^2+\sigma^2}, \tag{8}$$

When $\mu{>}\sigma$ and $\alpha^2{\ge}\frac{\sigma}{\rho^2}$, the equilibrium solution is

[13]For more information about pointwise algorithm, see the blog written by a Google engineer https://medium.com/@nikhilbd/pointwise-vs-pairwise-vs-listwise-learning-to-rank-80a8fe8fadfd.

$$s_i^* - s_j^* = x_L^* = \frac{-\sigma\mu\rho^2\alpha^2 - \rho^2\alpha c + \sigma^2\mu}{-\mu\rho^2\alpha^2 + \sigma^2}, \tag{9}$$

The condition $\alpha^2 < \frac{\sigma^2}{\rho^2 d_i}$ ensures that both the FOCs and the SOCs hold. The equilibrium point is a unique extremum and thus a global minimum.

To examine the effects of the API investment on the ranking stability, we let $P(\alpha)$ denote the probability that the phrase and/or sentence with the lowest distance has the lowest score, i.e., the company's ranking algorithm dominates. For the benchmark $\rho=1$, (1) when $\mu>\sigma$ and $\alpha^2 \geq \frac{\sigma}{\rho^2}$, we have $P^{'}(\alpha)>0$, (2) when $\mu<-\sigma$ and $\alpha^2 < \frac{\sigma}{\rho^2}$, we have $P^{'}(\alpha)<0$ and (3) when $0<\mu<\sigma$ and $\alpha^2 \geq \frac{\sigma}{\rho^2}$, we have $P^{'}(\alpha)<0$. For some $\sigma<1$ and an arbitrary ρ, (1) still holds. (2) and (3) hold for any σ and an arbitrary ρ.

To probe into the relationship between effectiveness U_i and API investment α, taking the derivative with respect to α of the effectiveness functions of both phrases and/or sentences shows that at the limit of $\alpha \to 0$, (1) when $0<\mu<\sigma$ and $\alpha^2 \geq \frac{\sigma}{\rho^2}$, or when $\mu<-\sigma$ and $\alpha^2 < \frac{\sigma}{\rho^2}$, the sign of $U^{'}(\alpha)$ is the opposite of the sign of μ, (2) when $\mu>\sigma$ and $\alpha^2 \geq \frac{\sigma}{\rho^2}$, since $U^{'}(\alpha)<0$, the effectiveness never increases for any of the equilibrium conditions.