Foundations and Trends® in Marketing

Leveraging Online Search Data as a Source of Marketing Insights

Suggested Citation: Rex Yuxing Du and Tsung-Yiou Hsieh (2023), "Leveraging Online Search Data as a Source of Marketing Insights", Foundations and Trends® in Marketing: Vol. 17, No. 4, pp 227–291. DOI: 10.1561/1700000070.

Rex Yuxing Du
University of Texas at Austin
rex.du@mccombs.utexas.edu

Tsung-Yiou Hsieh
Northeastern University
t.hsieh@northeastern.edu

now
the essence of knowledge
Boston — Delft

Contents

Leveraging Online Search Data as a Source of Marketing Insights

Rex Yuxing Du[1] and Tsung-Yiou Hsieh[2]

[1] *McCombs School of Business, University of Texas at Austin, USA; rex.du@mccombs.utexas.edu*
[2] *D'Amore McKim School of Business, Northeastern University, USA; t.hsieh@northeastern.edu*

ABSTRACT

Every year billions of users around the world submit trillions of queries through online search engines such as Google, Bing, Baidu, and Yandex. Over the years, aggregated and anonymized search volume data on keywords contained in all these queries have formed an epic database of human intentions that continues to expand every day. Thanks to platforms such as Google Trends, Google Ads Keyword Planner, Microsoft Advertising Keyword Planner, Baidu Index, and Yandex Wordstat, advertisers can readily assess search engine users' collective interests over time and across geographic areas to optimize their search engine marketing efforts. In this monograph, we illustrate how online search volume data, indexed or otherwise, can be leveraged as a powerful source of marketing insights for purposes beyond search engine marketing. We do so by offering a brief tutorial on Google Trends and Google Ads Keyword Planner, two popular (and free) platforms for gathering online search trend and volume data, respectively. We review prior studies that have examined the use of aggregate online search data

Rex Yuxing Du and Tsung-Yiou Hsieh (2023), "Leveraging Online Search Data as a Source of Marketing Insights", Foundations and Trends® in Marketing: Vol. 17, No. 4, pp 227–291. DOI: 10.1561/1700000070.

as (1) predictors for nowcasting and forecasting, (2) dependent variables in market response modeling, and (3) proxies for otherwise hard-to-measure constructs. In each of these three areas, we provide specific examples of applications to illustrate the power and versatility of online search data. We conclude by offering several ideas for future research where we see the full potential of online search data is still to be uncovered.

Keywords: online search; marketing insights; marketing research; big data.

1

Introduction

Online search has become an integral part of everyday life. From the mundane (e.g., the exact two-word phrase, "best toothbrush," is on average searched by twenty-two thousand Google users in the U.S. each month), to the more cerebral (e.g., each month, on average one-hundred-and-ten thousand Google users in the U.S. type into the query box the exact same question, "what is the meaning of life"), consumers try to resolve their queries with online search engines via their computers, mobile devices, and smart speakers, hoping that they can get what they are looking for in the returned results. Such a reliance on online search engines (e.g., Google, Bing, Baidu, and Yandex) is nowadays a routinized daily behavior for consumers around the world. For example, 84% of consumers conduct three or more searches per day through Google's search engine (Moz, 2022), which processes more than 8.5 billion queries every day (Internet Live Stats, 2022), amounting to more than 3.1 trillion searches a year.

As a byproduct of this process, aggregated and anonymized search volume data on keywords contained in all these queries have formed an epic database of human intentions that continues to expand every day. Thanks to platforms such as Google Trends, Google Ads Keyword

Planner, Microsoft Advertising Keyword Planner, Baidu Index, and Yandex Wordstat, advertisers can readily assess search engine users' collective interests over time and across geographic areas to optimize their search engine marketing efforts.

In this monograph, we draw on a growing literature that has illustrated how online search data, in indexes or volumes, can be leveraged as a powerful source of marketing insights for purposes beyond search engine marketing. For example, online search data has been used to investigate trends in consumer needs, wants, and preferences, or to assess consumer interests and concerns about different brands and products. As a source of marketing insights, online search data offers several advantages over survey and social media data.

First, what people type into search query boxes tends to be subject to less social disability biases, compared to how they respond to surveys or post on social media. Second, online search data, made available through platforms such as Google Trends, reflects the collective interests of the population by drawing on large and representative samples of the vast majority of online information seekers. As a result, online search data tends to be subject to less sample selection biases and sampling errors. Third, thanks again to platforms such as Google Trends, online search data can go as far back as 2004 and at the same time can be updated in near real time, providing an incredibly cost-effective way of gathering both historical and up-to-the-minute marketing insights, which can prove particularly valuable for marketers and researchers with limited resources. Consequently, we believe the availability of online search data has the potential of leveling the playing field when it comes to marketing intelligence.

Accordingly, our monograph is written with two main audiences in mind: practitioners seeking a supplemental data source for marketing insights and academics seeking an alternative data source for addressing their research questions. For practitioners, we aim to offer a guide on how best to utilize platforms such as Google Trends and extract actionable insights for a wide array of business decisions, illustrated with real-world example applications. For academics, we aim to provide a literature review and a framework that integrates the various avenues through which online search data can be leveraged in scientific research.

The rest of the monograph proceeds as follows. Section 2 offers a brief tutorial of Google Trends and Google Ads Keyword Planner, two popular platforms for gathering online search trends and volume data, respectively. We focus on lesser-known features and offer tips that we have found particularly useful in practice in order to get the most out of these platforms. We also briefly discuss Baidu Index as an alternative to Google Trends for insights about the Chinese market, where Baidu is the dominant search engine.

The next section offers a review of the literature that has utilized online search data. Section 3 surveys research that has treated aggregate online search interests as either concurrent or leading indicators of real-world phenomena (e.g., flu outbreaks, category demands, product sales, economic conditions). This stream of research focuses mainly on gauging the value of online search data as predictors in improving nowcasting or forecasting performances.

Section 4 examines research that has treated online search data as response variables that can help measure and improve marketing effectiveness in terms of both immediate and longer-term impacts.

Section 5 reviews research that has treated patterns of online searches as unvarnished reflections of the public psyche, uncovering what people really think, feel, and intend to do, insights that may otherwise be difficult to ascertain based on what people post on social media or tell market researchers in surveys.

Section 6 highlights several promising areas for future research where online search data can serve as a big-data supplement to traditional market research, e.g., integrating online search, social media, and survey data for better brand health tracking; using online search data to spot emergent trends in consumer needs and wants that can reshape market boundaries, while separating them from fleeting fads.

2

Tutorial on Gathering Online Search Data

With about 92% of the global search engine market as of January 2022, Google remains the unquestionable leader (GS Statcounter, 2022). In this section, we first provide a brief tutorial of Google Trends and Google Ads Keyword Planner, two popular (and free) platforms for gathering aggregated and anonymized online search trends and volume data, respectively. We focus on lesser-known features and offer tips that we have found particularly useful in practice in order to get the most out of these platforms. We also briefly introduce Baidu Index, an alternative platform for aggregate online search data for readers who are interested in the Chinese market, where Baidu is the leading search engine. We conclude this section with a discussion on how to identify keywords that are most relevant to the underlying constructs of interest, arguably the most important (and yet often neglected) step in gathering online search data.

2.1 Google Trends

The most popular source for online search data is probably Google Trends (GT for short hereinafter) https://trends.google.com/trends, which has been available to the public for free since 2008, then known

as Google Insights for Search. GT provides normalized volume indexes for queries people have been entering into Google's search engine, going all the way back to January 2004. Since 2016, GT data has been made available in real time (for the most recent seven days). The trillions of queries every year for nearly 20 years make GT one of the world's largest real-time databases. For more details about GT, see the official GT Help Center.[1]

To illustrate some of the features of GT that are particularly relevant for marketing research, we use the following example[2] (click on the link to see the exact settings), where we specify the GT user interface as follows:

- For the search term to enter into the query box, we type in two words "weather tomorrow" (without the double quotation marks for broad match);
- For the location, we set it to the United States;
- For the time range, we set it to the past 12 months;
- For the category, we set it to all categories;
- For the type of search, we set it to Google web search.

Given the above settings, GT returns a time plot of weekly indexes for the past 52 weeks (with each week starting on Sunday). The data behind the time plot can be downloaded in a CSV file. To generate the data, GT goes through the following steps. First, it draws a random sample of all the Google web searches (the type of search) conducted in the U.S. (the location) over the past 52 weeks (the time range). Second, within the sample, for each week, it calculates the percentage of searches that contain both "weather" and "tomorrow" regardless of the order or whether there are other words in between (the query). In other words, GT normalizes a particular week's volume of searches for the focal query by dividing it by that week's volume of all web searches. Lastly,

[1] https://support.google.com/trends#topic=6248052.
[2] https://trends.google.com/trends/explore?geo=US&q=weather%20tomorrow.

the percentages from the previous step are scaled proportionately on a range of 0 to 100.

Given the above data generating process, it is important to keep in mind that each data point on a GT time plot reflects search interest of the focal query relative to total searches from the location and time range represented by the data point. Furthermore, each data point contains a random sampling error, which can potentially be reduced by running the same GT query multiple times (thus triggering GT to draw multiple random samples) and taking the average.

In constructing the focal query, it is important to know how punctuations (e.g., the plus and minus signs, double quotation marks) work in GT to form composite queries, which is explained in detail on this GT Help Center page.[3] It is incumbent upon the researcher to construct a query that most accurately reflects the topic under study. For example, for search interest in "fuel efficiency" in car shopping, instead of using only the phrase "fuel efficiency," a more comprehensive composite query could be: "fuel efficiency" + "fuel economy" + "best gas mileage" + "best mpg" -motorcycle, where the double quotation marks indicate phrase match (i.e., the order of the words is fixed and there are no other words in between), the plus signs indicate an OR relationship, and the minus sign indicates excluding the term after it. We see that the composite query[4] has far greater search volume than the phrase "fuel efficiency" by itself.[5]

The user interface of GT allows comparisons of up to five queries at once. When multiple queries are entered, the maximum data point of all the queries over the entire selected time range is set to 100, and all the other data points are scaled proportionately. This way, the relative popularity of the search queries in comparison remains intact. In comparisons where more than five queries are involved, one can

[3] https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530.

[4] https://trends.google.com/trends/explore?geo=US&q=%22fuel%20efficiency%22%20%2B%20%22fuel%20economy%22%20%2B%20%22best%20gas%20mileage%22%20%2B%20best%20mpg%20-motorcycle,%22fuel%20efficiency%22.

[5] GT imposes a cap on the number characters allowed in each composite query, which appears to be less than 100.

apply the "transitivity rule" as a workaround for the five-query-per-comparison limit. For example, the first comparison includes queries A, B, C, D, and E, and the second comparison includes queries A, F, G, H, and I. From the first comparison, we know on average A is twice as popular as B, and from the second comparison, we know A is half as popular as G. The "transitivity rule" would indicate G is four times as popular as B and data from the second comparison can be rescaled accordingly to be directly comparable to data from the first comparison. The key to applying the "transitivity rule" is to make sure there is at least one overlapping query between two different comparisons.

Besides comparing different queries for the same location and time range, GT also allows one to compare

- The same query across different locations (see this GT example,[6] which compares "best mpg" searches between California and Texas over the past five years),
- Different query-location combinations (see this GT example,[7] which compares "best mpg" searches in California with "best gas mileage" searches in Texas over the past five years),
- Different query-location-time range combinations (see this GT example,[8] which compares "best mpg" searches in California in 2021 with "best gas mileage" searches in Texas in 2020).

This is accomplished by simultaneously changing the query and/or the location and time range filters for each query box. See this GT Help Center page[9] for more details on how to make comparisons between queries, locations, and time ranges.

[6] https://trends.google.com/trends/explore?date=today%205-y,today%205-y&geo=US-CA,US-TX&q=best%20mpg,best%20mpg.

[7] https://trends.google.com/trends/explore?date=today%205-y,today%205-y&geo=US-CA,US-TX&q=best%20mpg,best%20gas%20mileage.

[8] https://trends.google.com/trends/explore?date=2021-01-01%202021-12-31,2020-01-01%202020-12-31&geo=US-CA,US-TX&q=best%20mpg,best%20gas%20mileage.

[9] https://support.google.com/trends/answer/4359550?hl=en&ref_topic=4365530.

Through the GT user interface, one can set the location to a particular country or state/province. For the U.S., one can drill further down to a particular media market or Designated Market Area (DMA). When the location is set to a region where English is not the main language, the local language should be used in query construction. When multiple languages are used in a region, one can use the + sign to combine terms from different languages, including non-Latin characters. Of course, in the few countries where Google is not the dominant search engine (e.g., China, Russia), GT data is likely to be less representative.

The GT user interface allows one to set the time range in different ways, the most flexible of which would be "Custom time range," which offers two options: archive and past week. With "archive," one can set the "From" date and the "To" date. The "From" date can be as early as January 1, 2004 and the "To" date can be as recent as the present calendar day. Depending on the length of the specified time range, GT returns data of different granularity:

- When the time range covers about five years or more, the returned data is monthly;
- When the time range covers between nine months and five years, the returned data is weekly;
- When the time range covers about eight months or less, the returned data is daily.

With the "past week" option, GT returns real-time data with different levels of granularity depending on the length of the time range:

- When the time range covers between three and seven days, the returned data is hourly;
- When the time range covers between one and two days, GT returns data by 16-minute increments;
- When the time range covers between six and 24 hours, GT returns data by eight-minute increments;
- When the time range is five hours or less, GT returns data by one-minute increments.

An alternative to using the GT user interface in specifying a query with a custom time range is to enter a custom URL directly. For example, the link in the footnote[10] pulls minute-by-minute data from GT for a three-hour window during Super Bowl 2016, from 7PM to 10PM Central Standard Time (CST) on February 7, 2016.

The data pulled from the URL compares search interests for five queries: wix, buick, bud light, turbotax, and amazon, all brands that had an ad insertion during the Super Bowl 2016 broadcast. Note that, in the above URL, the time range is set to 2016-02-08T00 to 2016-02-08T03 in Coordinated Universal Time or UTC, which is five hours ahead of CST. In other words, to set a custom time range through a GT URL, one needs to use UTC.

By programming the starting and ending dates and times directly, one achieves maximum flexibility in pulling GT data for any time range. Of course, the granularity of the returned data still depends on the length of the time range. To pull more granular data over longer time ranges, one will need to use the "transitivity rule" to string together data pulled from multiple shorter time ranges and make sure the time ranges overlap partially for rescaling purposes (i.e., using the workaround discussed earlier when dealing with comparisons that include more than five queries).

In case a search query has multiple meanings, GT offers a category filter to refine the results. For example, for the query "apple," one can select a category to indicate whether one means the fruit (Food & Drink) or the maker of the iPhone (Computers & Electronics). In total, the category filter includes about two dozen categories, e.g., Arts & Entertainment, Autos & Vehicles, Real Estate, Travel. Within each category, there are multiple subcategories, e.g., Vehicle Maintenance, Vehicle Shopping under Autos & Vehicles. When a category or subcategory is selected and the query box is left empty, GT returns the search indexes for the selected category or subcategory, which can prove useful in measuring industry-wide search interests. That said, it is unclear how GT classifies searches into categories and how the underlying algorithm

[10] https://trends.google.com/trends/explore?date=2016-02-08T00%-202016-02-08T03&geo=US&q=wix,buick,bud%20light,turbotax,amazon.

changes over time. As a result, researchers need to use the category filter with caution and check face validity and robustness when the category filter is used.

Finally, besides web searches conducted on https://www.google.com/, GT also provides a filter for data on image searches conducted on https://images.google.com/, news searches on https://news.google.com/, shopping searches on https://shopping.google.com/, and video searches on https://www.youtube.com/. Researchers can use data for different types of searches creatively. For example, when the news search filter is selected, one often can see more distinct spikes in searches involving brand names. In this GT example,[11] the news searches for the brand "Chipotle" exhibit several large spikes between November 2015 and February 2016, a time period when the restaurant chain suffered through multiple food-related illness outbreaks, which generated intense media coverage and customer backlash. Because ordinary consumers usually do not search for news stories about a brand unless there is an unusual (and most often negative) event happening that is related to the brand (e.g., a product harm or public relations disaster). By applying the news search filter, one can potentially better measure and compare the relative magnitudes of different brand crises. For example, "Is the current crisis generating more or less consumer interest than the previous one, as manifested in online news searches?"

In summary, through the above tutorial, we have attempted to highlight a few features of GT that we believe can prove particularly useful in marketing research. As a data source, GT has many merits:

- **Accessibility** – available to the public free of charge, downloadable, reproducible (through custom URLs), and available in real time or near real time.
- **Comprehensiveness** – covering all the searches conducted via Google, the predominant platform for online searches in most countries around the world.

[11] https://trends.google.com/trends/explore?date=2014-01-01%202018-12-31&geo=US&gprop=news&q=chipotle.

- **Temporal granularity** – available by month, week, day, hour, or minute.
- **Geographical granularity** – available by country, state/province, or city/DMA.
- **Historical coverage** – going as far back as 2004.
- **Diversity** – covering searches conducted in different languages.
- **Scalability** – with programmable URLs, codes can be deployed to pull data automatically.[12]

2.2 Google Ads Keyword Planner

The many merits of GT aside, it does have several limitations as a source for online search data. First, GT only provides search trends in indexed values. Although the "transitivity rule" allows one to obtain the relative magnitudes of multiple search trends, one still does not know their absolute magnitudes, which can be important in some applications. Second, GT can run into data sparsity issues for queries that are not searched frequently enough in a particular geographic area (e.g., an uncommon query in a small DMA), resulting in missing or unreliable search trend data. Third, it becomes cumbersome to pull data from GT when a large number of search queries is involved.

Fortunately, Google provides another tool for pulling actual search volume data: Google Ads Keyword Planner (GAKP for short hereinafter). To access GAKP, one needs to have a Google Ads account, which can be created for free (see https://ads.google.com/home for instructions). To have access to the full functionality of GAKP (e.g., exact search volume data as opposed to broad ranges such as 100–1K or 1–10K), one needs to set up an active ad campaign in their Google Ads account (for instructions on how to do so see https://support.google.com/google-ads/answer/6324971?hl=en).

[12]For example, R package "gtrendsR" or Python package "pytrends" downloads data from GT automatically.

After logging into the Google Ads account and navigating to GAKP, one will see two different tools: "Discover new keywords" and "Get search volume and forecasts." The former is mainly used to generate related keywords given a list of seed keywords (up to ten) or a seed website or webpage. For example, one can use the name of a brand or product and its common variants as seed keywords and use the "Discover new keywords" tool to generate keywords that include terms that are frequently co-searched with the focal brand or product (e.g., competitors' brand names, product attributes).

The "Discover new keywords" tool is particularly useful in contexts where search engine users may type in all sorts of queries when seeking information on the topic of interest to the researcher. For example, when "fuel efficiency" is the topic and used as the seed keyword, the top related keywords suggested by the "Discover new keywords" tool (out of a list of more than 1,200) include: "best gas mileage suv," "best gas mileage cars," "best mpg cars," "suv with best mpg," "fuel economy," "most fuel-efficient cars," etc. It soon becomes clear that consumers use queries such as "best gas mileage," "best mpg," and "fuel economy" as alternatives to "fuel efficiency." One can use these newly discovered keywords as seeds to identify additional related keywords and repeat the process until no new relevant keywords appear.

In addition to seed keywords entered by the researcher, the "Discover new keywords" tool also recommends additional seed keywords through the "Broaden your search" feature on the results page. In short, the "Discover new keywords" tool can be used interactively and iteratively to generate a comprehensive list of keywords that are relevant to the topic under study, including the long-tail ones that may otherwise be ignored by the researcher.

Often times the "Discover new keywords" tool may return hundreds or even thousands of keywords Google deems as relevant. On the GAKP results page, one can use the "Add filter" feature to shorten the list. For example, the "Keyword" filter allows one to zero in on all the keywords that contain or do not contain certain terms. The "Avg. monthly searches" filter allows one to focus on keywords with search volumes that are above or below certain thresholds.

When one has already got a comprehensive list of the relevant keywords, they can copy and paste or upload the list to the "Get search volume and forecasts" tool of GAKP, which would return monthly search volumes for each keyword on the list. The search volumes provided by GAKP are based on exact match, as opposed to phrase match or broad match in GT.[13] For example, the search volume for "basketball hoop" does not include searches for "kids basketball hoop" or "portable basketball hoop." In other words, to estimate the volume of searches containing the phrase "basketball hoop," one needs to have a comprehensive list of all the common search keywords that contain the phrase "basketball hoop" and sum up the search volumes given by GAKP for each keyword on the list.

GAKP also offers several filters for customizing the returned search volume data. First, the "Location" filter allows one to specify the geographic area under study, which can be a country, province/state, city/DMA, county, etc. Up to ten locations can be selected at the same time. Second, the "Language" filter allows one to specify the language of interest to the researcher, which can prove handy when conducting international marketing research. Third, the "Search network" filter allows one to choose between "Google" (by default) and "Google and search partners." We recommend the latter in most applications because it would include searches conducted on hundreds of non-Google websites, as well as YouTube and other Google platforms such as Google Shopping, Google Maps, Google Images. As researchers, we care about the search per se, as opposed to on which website or platform it is conducted. Finally, the "Date range" filter allows one to customize the time window. The default is the most recent twelve months, and the maximum is the most recent 48 months, which can be a limiting factor when data going farther back is needed.

Finally, unlike GT search trend index data, which is available in real or near real time, GAKP search volume data is only available by month and is updated about two to three weeks after the end of a calendar

[13]For certain close variants, e.g., best mpg car vs. best mpg cars, GAKP treats them as the same keyword and reports the combined search volume.

Table 2.1: Google Trends vs. Google Ads Keyword Planner

Feature	Google Trends	Google Ads Keyword Planner
Access	Free	Require a Google Ads account (free) and running an active ad campaign (for full functionality)
Data type	Normalized index indicating share of all searches	Search volume
Data sampling	Random sampling	Census
Data sparsity	Can be an issue for infrequently searched queries or small geographic areas	Not an issue
Timeliness	Real time/near real time	With a lag of two to three weeks
Time range	2004 through present	Most recent 48 months
Time granularity	Month, week, day, hour, minute	Month
Composite query	Allowed	N/A
Match type	Phrase match and broad match	Exact match
Category filter	Yes	N/A
Search type filter	Web, news, image, shopping, YouTube	Google, Google and search partners
Discovering related keywords	Limited	Comprehensive
Programmable URL	Yes	N/A
Max # of queries submitted at once	Five	Thousands

month. Table 2.1 summarizes the differences between GT and GAKP as sources for online search data.

2.3 Baidu Index

China is one of the few markets where Googles is not the dominant search engine. Baidu accounts for over 75% of the Chinese search engine market. Like GT, Baidu Index https://index.baidu.com/ (BI for short hereinafter) also allows users to gather data on the popularity and trends of keywords searched on Baidu. BI provides data on search volume, geographic distribution, related queries, and other relevant information via an interface that is very similar to GT's. That said, there are a few notable differences between BI and GT.

First, instead of providing normalized search volume indexes, BI provides actual search volume data. Second, since most people who search on Baidu enter their queries in Mandarin, users of BI need to construct their queries accordingly. Third, BI provides search volume data based on "exact match," without the option of "broad match." Last, in addition to regions, BI also allows users to break down search volume data by demographics such as age and gender. For more details about BI, we refer interested readers to Vaughan and Chen (2014), who provide an excellent comparison of BI and GT.

2.4 Identifying Relevant Keywords for Online Search Data Collection

In our experience, the most important (and yet often neglected) step in collecting online search data lies in the identification of keywords that are most relevant to the underlying constructs of interest. This can be particularly challenging when information seekers with the same intent can use any number and combination of keywords in their search queries. An inclusion of irrelevant ones and/or an exclusion of relevant ones can add biases as well as noises to the resulting data and undermine the power and validity of subsequent analyses.

Often times researchers have relied primarily on their intuition and domain knowledge in identifying relevant keywords. A few have attempted to be less ad hoc. For example, from a pool of 50 million search queries, Ginsberg *et al.* (2009) identify 45 that exhibit the highest correlation with the variable they try to predict (influenza-like illness physician visits). Brynjolfsson *et al.* (2016) show that relevant keywords may be identified through crowdsourcing by asking survey respondents to perform word association tasks with each focal construct.

Du *et al.* (2015) illustrate how GAKP's "Discover new keywords" tool can be leveraged in an attempt to be more systematical in identifying relevant keywords. Their approach consists of the following steps:

1. Identify keywords that are commonly associated with the focal constructs (in their case, consumer interests in various vehicle features) by examining sources of user generated content such as online product reviews;

2. Enter the keywords identified in (1) into GAKP's "Discover new keywords" tool for additional recommendations and select the high-frequency ones with the "Avg. monthly searches" filter;[14]

3. Repeat (2) using the newly identified keywords as seeds and iterate until no new high-frequency relevant keywords emerge;

4. Construct a composite query using all the high-frequency relevant keywords identified.

[14] As a supplemental source, GT has a "Related queries" feature that can also provide additional relevant keywords.

3

Online Search Data as Predictors

In this section, we review research that has treated aggregate online searches as either concurrent or leading indicators of real-world phenomena and used search index or volume data for nowcasting ("predicting the present") or forecasting ("predicting the future"). Many studies have investigated whether and how online search data can be leveraged as additional predictors in nowcasting or forecasting models, with the incremental predictive power of online search data coming mainly from two sources.

First, predictors based on online search data are often available with less time delay than predictors from other sources. For example, economic indicators such as gross domestic product, unemployment rate, and consumer price index are typically released by government agencies with a reporting lag of several weeks, many of which are often revised a few months later. Similarly, firms' key performance indicators such as market shares and survey-based brand health measures also tend to have a substantial reporting lag. In contrast, GT data is available in real or near real time (e.g., by the next minute, hour, day, week, or month depending on the desired level of granularity).

Researchers can take advantage of the timeliness of online search data to improve both nowcasting and forecasting. For example, the Mortgage Bankers Association in the U.S. publishes weekly mortgage application

indexes based on surveys of mortgage bankers, which have a reporting lag of two weeks. GT weekly indexes for queries such as "mortgage rates" and "mortgage calculator" are available with no reporting lag and thus can be used for nowcasting mortgage application activities.

Second, online searches are often conducted in the early stages of certain processes (e.g., a purchase funnel consisting of awareness, interest, consideration, preference, and purchase), thus providing leading indicators of the behavioral outcomes under study (e.g., product sales). Under such circumstances, the incremental predictive power of online search data depends on (a) the amount of lead time between online searches and the behavioral outcome of interest, and (b) the stability of conversion from online searches to the behavioral outcome of interest. Both the lead time and the stability of conversion are empirical questions. For example, queries such as "mortgage help" and "hardship letter" are searched by financially distressed households when they are initially confronted with the risk of mortgage default, a process that can take months to unfold. As a result, researchers can include online searches for queries such as "mortgage help" and "hardship letter" as leading indicators in forecasting actual mortgage defaults.

3.1 Review of Existing Studies

Among the numerous studies utilizing online search data as predictors for nowcasting or forecasting, Ginsberg *et al.* (2009) and Choi and Varian (2012) are two of the best known and most cited. Ginsberg *et al.* (2009) develop an automated method for selecting search queries related to the spread of influenza and build a nowcasting model for the prevalence of influenza-like illness (ILI) physician visits. They show that their weekly nowcasts based on Google search data are highly correlated with the ILI levels reported by the Centers for Disease Control and Prevention (CDC), obtaining a mean correlation of 0.97 in one holdout test. The main advantage of their method lies in that their nowcasts are made one to two weeks ahead of the CDC reports, offering an earlier detection of influenza outbreaks and thus allowing the healthcare system to respond more rapidly. Another advantage of their online search data-based nowcasting method is that it can be applied to more

granular geographic areas. For example, while the CDC does not make state-level data publicly available, data from GT can be drilled down to metropolitan areas.

Choi and Varian (2012) demonstrate how GT data can be used to improve near-term forecasts of economic indicators such as retail sales, automobile sales, home sales, and travel. Their findings suggest that online search data-based predictors can improve predictive performances by 9.3% to 13.6%.

Inspired by these two pioneering studies, researchers have gone on to show that predictors derived from online search data can improve nowcasting and forecasting performances in various contexts. Besides GT, similar data from Yahoo!, Baidu, MSN, and Wikipedia have also been used in prior studies, reporting out-of-sample improvements in predictive performance ranging from as low as 4% to as high as 81%. Table 3.1 summarizes some of the more cited studies in this stream of research, covering a wide array of fields:

- **Epidemiology** (Chan *et al.*, 2011; Dugas *et al.*, 2012; Ginsberg *et al.*, 2009; Pelat *et al.*, 2009; Polgreen *et al.*, 2008; Santillana *et al.*, 2015; Seifter *et al.*, 2010; Teng *et al.*, 2017; Yang *et al.*, 2015a).

- **Economics and Social Sciences** (Askitas and Zimmermann, 2009; Brynjolfsson *et al.*, 2016; Choi and Varian, 2009, 2012; D'Amuri and Marcucci, 2017; Ettredge *et al.*, 2005; Goel *et al.*, 2010; Vosen and Schmidt, 2011; Wu and Brynjolfsson, 2009; Yu *et al.*, 2019).

- **Tourism** (Bangwayo-Skeete and Skeete, 2015; Li *et al.*, 2017; Yang *et al.*, 2015b).

- **Finance** (Bijl *et al.*, 2016; Da *et al.*, 2011; Dimpfl and Jank, 2016; Kristoufek, 2013; Preis *et al.*, 2010, 2013).

- **Marketing** (Du *et al.*, 2015; Hu *et al.*, 2014; Kulkarni *et al.*, 2012; Xiong and Bharadwaj, 2014).

Table 3.1: Studies using online search data as predictors in nowcasting or forecasting

Paper	Topic Domain	DV	Aim of Study	Time Period	Search Data Source	Methodology	Search Term Selection	Improvement
Ettredge *et al.* (2005)	Economics	Unemployment	Forecasting	2001–2003	WordTracker's Top 500 Keyword Report	Linear Regression	Self Defined Keywords	N/A
Polgreen *et al.* (2008)	Epidemiology	Weekly Influenza Cultures and Mortality	Nowcasting, Forecasting	2004–2008	Yahoo	Linear Regression	Self Defined Keywords	N/A
Askitas and Zimmermann (2009)	Economics	Unemployment	Nowcasting	2004–2009	Google Insights	Linear Regression	Self Defined Keywords	N/A
Choi and Varian (2009)	Economics	Initial Claims for Unemployment Benefits	Nowcasting	2004–2009	Google Trends	ARIMA	Related Google Category	12.9∼15.74%
Wu and Brynjolfsson (2009)	Economics	Home Sales	Nowcasting	2007-2009	Google Trends	AR Model	Related Google Category	76%
Ginsberg *et al.* (2009)	Epidemiology	ILI Physician Visits	Nowcasting	2004–2008	Google Trends	Linear Regression	Correlation-Based Method	N/A
Pelat *et al.* (2009)	Epidemiology	ILI Incidence Rate	Correlational	2004–2009	Google Trends	–	Self Defined Keywords	N/A
Seifter *et al.* (2010)	Epidemiology	Lyme Disease	Correlational	2003–2009	Google Trends	–	Self Defined Keywords	N/A

Continued.

Table 3.1: Continued.

Paper	Topic Domain	DV	Aim of Study	Time Period	Search Data Source	Methodology	Search Term Selection	Improvement
Preis *et al.* (2013)	Finance	Transaction Volume of S&P 500	Correlational	2004–2010	Google Trends	–	Company Name	N/A
Goel *et al.* (2010)	Social science	Box Office Revenue, Video Game Sales, Billboard Top Songs	Forecasting	2008–2009	Yahoo!	Linear Regression	Self Defined Keywords	3.75%–81%
Vosen and Schmidt (2011)	Economics	Private Consumption	Nowcasting, Forecasting	2005-2009	Google Trends	AR Model	Related Google Category	21% for Nowcast, 35% for Forecasting
Chan *et al.* (2011)	Epidemiology	Dengue Cases	Nowcasting	2003–2010	Google Trends	Linear Regression	Self Defined Keywords	N/A
Da *et al.* (2011)	Finance	Stock Return	Nowcasting	2004–2008	Google Trends	VAR	Stock T icker	N/A
Choi and Varian (2012)	Economics	Automotive Sales, Unemployment Claims, Travel Planning, and Consumer Confidence	Nowcasting	2004–2011	Google Trends	AR Model	Related Google Category	9.3%∼13.6%
Dugas *et al.* (2012)	Epidemiology	ILI Cases	Correlational	2009–2010	Google Trends	–	N/A	N/A

Continued.

Table 3.1: Continued.

Paper	Topic Domain	DV	Aim of Study	Time Period	Search Data Source	Methodology	Search Term Selection	Improvement
Kulkarni *et al.* (2012)	Marketing	New Product Sales	Forecasting	2006	Google/Yahoo!/ MSN	Probablistic Model	Movie Titles	N/A
Kristoufek (2013)	Finance	BitcoinPprice	Forecasting	2011-2013	Google Trends, Wikipedia	VAR and VECM	Self Defined Keywords	N/A
Preis *et al.* (2013)	Finance	Dow Jones Industrial Average	Correlational	2004–2011	Google Trends	–	Self Defined Keywords	N/A
Hu *et al.* (2014)	Marketing	Automobile Sales	Forecasting	2004-2012	Google Trends	State Space Model	Heuristic Method (Keyword Planner)	4.5%–15.7%
Xiong and Bharadwaj (2014)	Marketing	Video Game Sales	Forecasting	2009–2010	Google Trends	Functional Data Analysis	Self Defined Keywords	28.1%–35.8%
Yang *et al.* (2015a)	Epidemiology	ILI Cases	Nowcasting	2009–2015	Google Trends, Google Correlate	AR Model	Self Defined Keywords, Google Correlate	35.10%
Santillana *et al.* (2015)	Epidemiology	ILI Cases	Nowcasting, Forecasting	2011–2015	Google Trends	Machine Learning	Self Defined Keywords	53% for Nowcast, 48% for Forecasting
Yang *et al.* (2015a)	Tourism	Visiting	Forecasting	2006–2013	Google Trends, Baidu Index	ARMAX	Correlation-Based Method	36.8%–80.1%

Continued.

Table 3.1: Continued.

Paper	Topic Domain	DV	Aim of Study	Time Period	Search Data Source	Methodology	Search Term Selection	Improvement
Bangwayo-Skeete and Skeete (2015)	Tourism	Visiting	Forecasting	2004–2012	Google Trends	Mixed-Data Sampling Model	Self Defined Keywords	N/A
Du *et al.* (2015)	Marketing	Automobile Sales	Forecasting	2004–2011	Google Trends	Hierarchical Model	Self Defined Keywords	10.54%
Brynjolfsson *et al.* (2016)	Economics	ILI Cases and Unemployment Benefit Claims	Forecasting	2004–2011	Google Trends	AR Model	Crowd-Squared Method	11.90%
Dimpfl and Jank (2016)	Finance	Trading Volume	Forecasting	2006-2011	Google Trends	VAR	Self Defined Keywords	4.50%
Bijl *et al.* (2016)	Finance	Stock Return	Forecasting	2008-2013	Google Trends	Linear Regression	Company Name	N/A
D'Amuri and Marcucci (2017)	Economics	Unemployment	Nowcasting, Forecasting	2004–2008	Google Trends	AR Model	Self Defined Keywords	44% for Nowcast, 17% for Forecasting
Teng *et al.* (2017)	Epidemiology	Zika Virus Cases	Forecasting	2011-2015	Google Trends	ARIMA	Self Defined Keywords	N/A
Li *et al.* (2017)	Tourism	Visiting	Forecasting	2011–2015	Baidu Index	Generalized Dynamic Factor Model (GDFM)	Self Defined Keywords	22%
Yu *et al.* (2019)	Economics	Oil Consumption	Forecasting	2004–2015	Google Trends	Econometric Model and Machine Learning	Self Defined Keywords	1.40%

In the marketing literature, researchers have used online search data mainly as indicators of consumer interests in the focal products or brands under study, which are treated as noisy but nevertheless informative signals of purchase intentions and thus can be used to better predict product sales. For example, Kulkarni *et al.* (2012) build a model that leverages online search data in forecasting opening-week sales of movies. Xiong and Bharadwaj (2014) develop a functional data analysis method to predict video game sales as a function of prerelease search patterns. Both studies find that prerelease product searches are predictive of post-release product sales.

Going beyond online searches for products or brands, Du *et al.* (2015) develop a model that leverages online searches for product features (e.g., fuel efficiency) in predicting product sales (e.g., Prius). They find that product feature searches are positively correlated with product feature importance, and the predictive performance of market response models can be improved substantially by augmenting marketing mix data and product search data with product feature search data.

Besides identifying the most relevant search queries as predictors in nowcasting or forecasting, researchers also need to select the modeling framework that can best leverage the signals contained in those predictors. While most of the existing studies in this area have relied on time series models (e.g., Auto Regressive, Vector Auto Regressive, ARIMA, Dynamic Linear Model), a few have applied machine learning methods such as support vector machines (SVM), neural networks, or tree-based models (e.g., Yu *et al.*, 2019). Santillana *et al.* (2015) find that three machine learning methods—Lasso regression, SVM, and AdaBoost regression trees—outperform benchmark autoregressive models when online search data-based predictors are included. This suggests that, to improve predictive performance, marketing researchers need to explore a wide array of options in both the search queries used in constructing the predictors and the modeling methods for leveraging those predictors.

3.2 Challenges with Using Online Search Data as Predictors

Improved predictive performances reported in the literature aside, Lazer *et al.* (2014) raise two critical issues that could diminish the power of online search data as predictors in real-world applications: big data hubris and algorithm dynamics. The former refers to the often-implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Researchers need to be mindful that online search data are not initially designed for the nowcasting or forecasting tasks at hand, and mining predictors from a large pool of search queries can result in overfitting, which happens when the chosen predictors and model become too tailored to existing data and lose applicability to newer data. For example, the predictors that Ginsberg *et al.* (2009) use to nowcast the prevalence of influenza-like illness physician visits are selected from a pool of 50 million common search queries, a situation where the risk of overfitting is paramount. Indeed, subsequent real-world applications of their model faced substantial deterioration in predictive performance (Lazer *et al.*, 2014).

The other challenge with using online search data as predictors is algorithm dynamics. Online search engines continuously update their algorithms. As a result, how people search online or how search data are collected changes over time. Such unknown shifts in the underlying data generating process can negatively impact the accuracy of nowcasting and forecasting models using online search data as predictors. For example, Choi and Varian (2012) use category search indexes defined by GT as additional predictors in their models. Unfortunately, how GT groups search queries into different categories is a black box to researchers. When GT decides to update its definition of these category search indexes, it is questionable whether the same level of predictive performance will hold. Consequently, researchers need to be cognizant of the reality that the shelf life of their model using online search data can be quite limited, and their model requires close monitoring and may need frequent recalibration periodically.

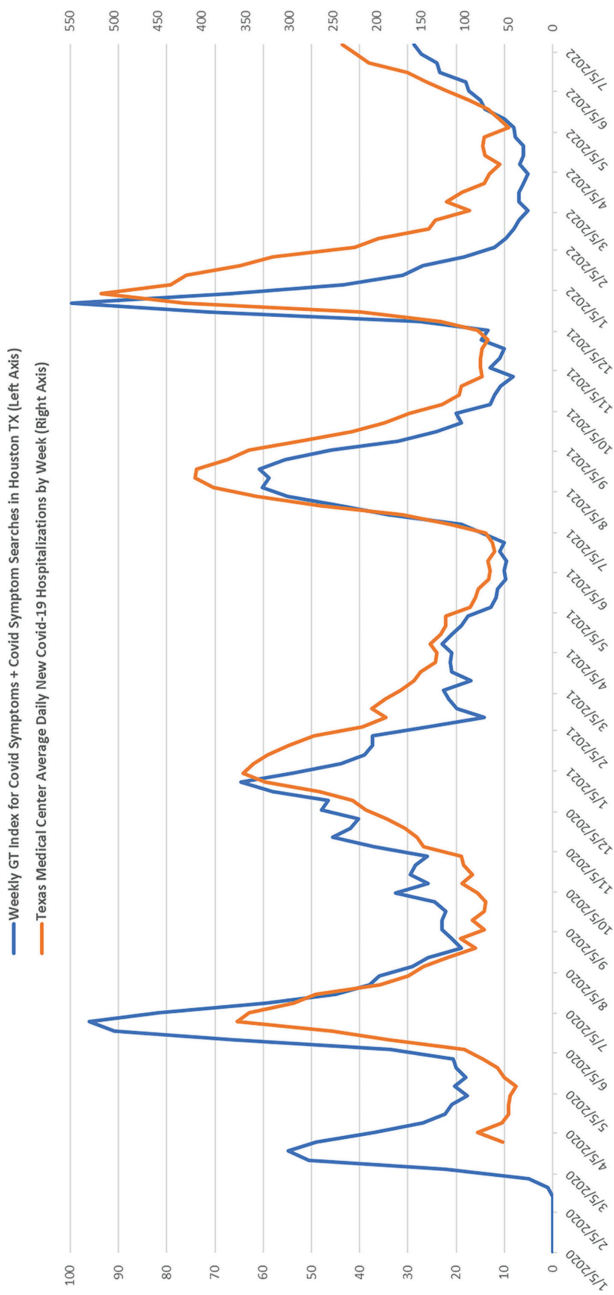

Figure 3.1: Online searches of Covid symptoms vs. new Covid-19 hospitalizations in Houston TX.

3.3 An Application of Using Online Search Data for Nowcasting

In the rest of this section, we illustrate the use of online search data for nowcasting with one specific application. Figure 3.1 presents two weekly time series: GT indexes for the composite query "covid symptoms+covid symptom" in the Houston DMA (plotted against the left y-axis) and average daily new Covid-19 hospitalizations (Monday through Sunday) in the Texas Medical Center (TMC), a coalition of major hospital systems in the Greater Houston Area (plotted against the right y-axis).[1]

The correlations between new Covid-19 hospitalizations and Covid symptom searches of zero, one, two, three, four, and five weeks of lead are, respectively, 0.72, 0.79, 0.73, 0.61, 0.45, and 0.29. This suggests that Covid symptom searches are likely an indicator of new Covid-19 hospitalizations with an average lead time of about one week, which is intuitive because there could be, on average, a one-week delay between the onset of Covid-like symptoms, which prompts people to seek information online, and the need for medical care in a local hospital. Moreover, the new Covid-19 hospitalization data is available with a one-week delay because TMC has to wait for reports from member hospitals. In contrast, GT data is available without delay. As a result, this creates a situation where online search data can potentially be used as predictors to nowcast a high-stake, real-world variable.

To quantify the dynamics between these two time series, we use the following Unobserved Components Model (Harvey, 1989):

$$y_t = \beta_t^0 x_t + \beta_t^1 x_{t-1} + \beta_t^2 x_{t-2} + e_t \tag{3.1}$$
$$e_t = \rho e_{t-1} + v_t \tag{3.2}$$
$$\beta_t^0 = \beta_{t-1}^0 + \varepsilon_t^0 \tag{3.3}$$
$$\beta_t^1 = \beta_{t-1}^1 + \varepsilon_t^1 \tag{3.4}$$
$$\beta_t^2 = \beta_{t-1}^2 + \varepsilon_t^2 \tag{3.5}$$

[1] GT data can be retrieved from https://trends.google.com/trends/explore?date=2020-01-01%202022-07-16&geo=US-TX-618&q=covid%20symptoms%20%2B%20covid%20symptom; hospitalization data can be retrieved from https://www.tmc.edu/coronavirus-updates/average-daily-new-covid-19-hospitalizations-by-week-monday-sunday/.

where y_t denotes the average daily new Covid-19 hospitalizations in week t; x_t denotes the GT index for Covid symptom searches in week t; e_t denotes noise in the data generating process, which is assumed to be autoregressive with a damping factor of ρ and a random disturbance $e(t)$ that is distributed i.i.d. normal with mean 0 and standard deviation of σ_e; ε_t^0, ε_t^1, and ε_t^2 denote shocks to the regression coefficients that capture the evolving conversion dynamics between Covid symptom searches and new Covid-19 hospitalizations, which are assumed to be distributed i.i.d. normal with mean 0 and standard deviation of σ_ε. β_t^0, β_t^1, and β_t^2 are the time-varying latent state variables to be inferred given the observed data and model parameters ρ, σ_y, and σ_ε, which are estimated using the SAS procedure UCM.[2] Based on data from the week of March 29, 2020 through the week of July 10, 2022, the calibrated model produces an R-square of 0.953 and a Random Walk R-square of 0.651 (i.e., capturing 65.1% of the variance of the residuals of a random walk model).

Figure 3.2 plots the model-inferred β_t^0, β_t^1, and β_t^2 over the 120-week observation window. We see that the rate at which Covid symptom searches convert to new Covid-19 hospitalizations varies depending on the time lag and evolves across different waves of the pandemic. In 2022, during the last wave in the observation window, one unit increase in Covid symptom search GT index converts to about 2.5 more daily new Covid-19 hospitalizations in the same week (β_t^0), between 3 and 3.5 more daily new Covid-19 hospitalizations one week later (β_t^1), and 2 more daily new Covid-19 hospitalizations two weeks later (β_t^2).

To assess the model's predictive performance, we simulate the following nowcasting task. First, we estimate the model using data through week 16 of the 120-week observation window, i.e., all the x_t's and y_t's observed through the week of July 12, 2020. Then, given x_{17}, we nowcast y_{17} using the calibrated model. We recalibrate the model using data observed through week 17 and then nowcast y_{18} given the updated model and x_{18}. The process is repeated through week 120. Figure 3.3 plots these one-step-ahead out-of-sample nowcasts against the actual

[2]Model specifications with fewer or more lags are rejected due to lower adjusted R-square. The model assumes an intercept of zero because when there is no Covid symptom searches, there should be, in all likelihood, negligible new Covid-19 hospitalizations. R package rucm is an alternative to SAS procedure UCM.

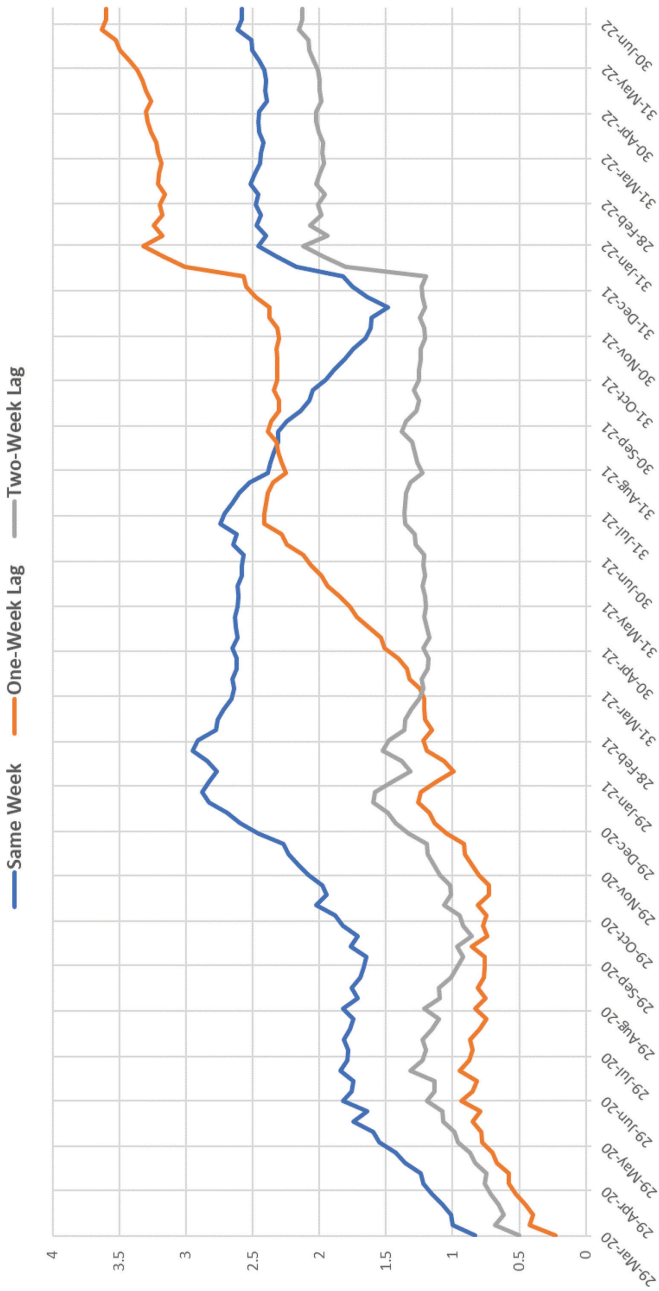

Figure 3.2: Rate of Covid symptom searches converting to daily new Covid-19 hospitalizations in Houston TX.

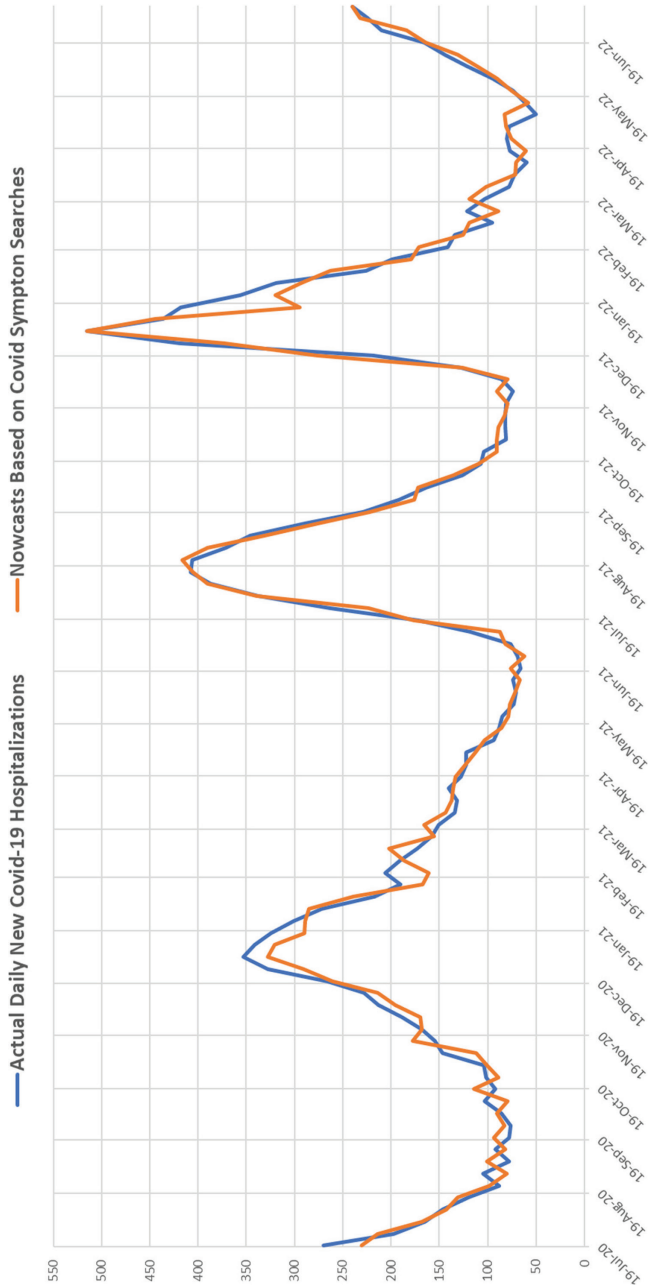

Figure 3.3: Goodness of fit of out-of-sample one-step-ahead nowcasts based on online search data.

daily new Covid-19 hospitalizations for week 17 through week 120. The 104 nowcasts have a Mean Absolute Percentage Error (MAPE) of 10.1%, an R-square of 0.956, and a Random-Walk R-square of 0.664.[3] Besides the remarkable goodness-of-fit, we see that the nowcasts accurately capture the inflection points of each wave, illustrating the potential power of online search data in predicting real-world phenomena when the relevant search queries and flexible modeling methods are used.

[3]In comparison, an AR(1) model of y_t produces an in-sample MAPE of 15.5%, an R-square of 0.862, and a Random-Walk R-square of -0.032, which indicates that an AR(1) model is no more accurate than a random walk model.

4

Online Search Data as Response Variables

4.1 Review of Existing Studies

In this section, we review studies in the marketing literature that have treated aggregate online searches as response variables that can help measure and improve marketing effectiveness in terms of both immediate and longer-term impacts. Table 4.1 provides an overview of this stream of research.

A growing body of research has paid particular attention to the causal linkage between TV advertising and online search as the so-called second-screen phenomenon, which refers to the use of an additional electronic device while watching TV, has become increasingly common. Zigmond and Stipp (2010) report the first case studies demonstrating that Google searches for the focal brands spike immediately after their TV ads during the opening ceremonies of the 2008 and 2010 Olympic Games. They show the potential of using online search data as an outcome measure in evaluating the causal impact of TV advertising. Since their pioneering work, many studies have examined how advertising drives online search for the focal brands using monthly, weekly, daily, hourly, or minute-level data in different product categories (e.g., Chandrasekaran *et al.*, 2018; Du *et al.*, 2019; Guitart and Stremersch,

Table 4.1: Studies using online search data as dependent variables in market response modeling

Paper	Advertising Variable	Aim of Study	Product Category	Time Period	Time Window	Search Data Source	Ad Content	Methodology	Ad Elasticity
Zigmond and Stipp (2010)	Ad Airing	Correlational	Multiple	2009	Minute	Google T rends	No	Visualization	–
Laroche *et al.* (2013)	Ad Expenditure	Correlational	Telecom	2007–2008	Weekly	Google, Yahoo!, MSN	No	VARX	0.008
Reiley and Lewis (2013)	Ad Airing	Correlational	Multiple	2011	Minute	Yahoo!	No	T-Test	–
Hu *et al.* (2014)	Ad Expenditure	Causal Inference	Automobile	2004–2012	Monthly	Google	No	State Space Model	0.03
Joo *et al.* (2014)	Ad Expenditure	Causal Inference	Financial Service	2011	Hourly	Google Trends	No	Log-Linear Model	0.17
Joo *et al.* (2016)	Ad Expenditure	Causal Inference	Financial Service	2006	Hourly	AOL	Yes	Choice Model	–
Chandrasekaran and Tellis (2007)	Ad Airing	Causal Inference	Multiple	2004–2012	Three-Day Window	Google Trends	Yes	Log-Linear Model	–
Du *et al.* (2019)	Audience Size	Causal Inference	Automobile	2015–2016	Minute	Google Trends	Yes	Non-Linear Regression	0.09
Guitart and Stremersch (2020)	Ad Expenditure	Causal Inference	Automobile	2004–2010	Monthly	Google Trends	Yes	Log-Linear Model	0.032

2020; Hu *et al.*, 2014; Joo *et al.*, 2014, 2016; Laroche *et al.*, 2013; Reiley and Lewis, 2013).

Using weekly data, Laroche *et al.* (2013) investigate the relationship between multi-channel advertising and consumer online searches for a telecommunication brand. Their findings show that exposures to advertising on different media outlets increase consumers' follow-up searches. With hourly data, Joo *et al.* (2014) examine the relationship between TV advertising and consumer online search behavior in the financial service category. They find that TV advertising increases the number of related Google searches and searchers' tendency to use branded instead of generic keywords. Using minute-level data, Du *et al.* (2019) quantify the immediate impact of TV advertising on brand search and price search in the automobile market, providing a framework for advertisers to enrich their media planning and campaign evaluations using highly granular online search data.

A few studies in this area have investigated the role of ad content in moderating the impact of advertising on online searches for the advertised brands. For example, using a sample of Super Bowl ads and online brand search data, Chandrasekaran *et al.* (2018) examine the moderation effects of four aspects of TV ad content: informational, emotional, prior media publicity, and brand website prominence in the ad. They find that informational ad content significantly increases online brand search while emotional ad content does not. Similarly, Guitart and Stremersch (2020) investigate the effects of informational and emotional ad contents on sales and online search. They show that while both emotional and informational ad contents increase sales, only emotional ad content increases online search. Du *et al.* (2019) also examine the moderation effects of different types of ad content. They report that ad creatives that are more likable, informative, and desirable tend to generate more post-ad online brand searches.

With monthly data, Hu *et al.* (2014) examine the impact of advertising on generating pre-purchase information search and converting information seekers into purchasers in the automobile market. They propose a dynamic linear model that decomposes the overall impact of advertising into a component that influences online search and another component that influences the conversion from online searches to

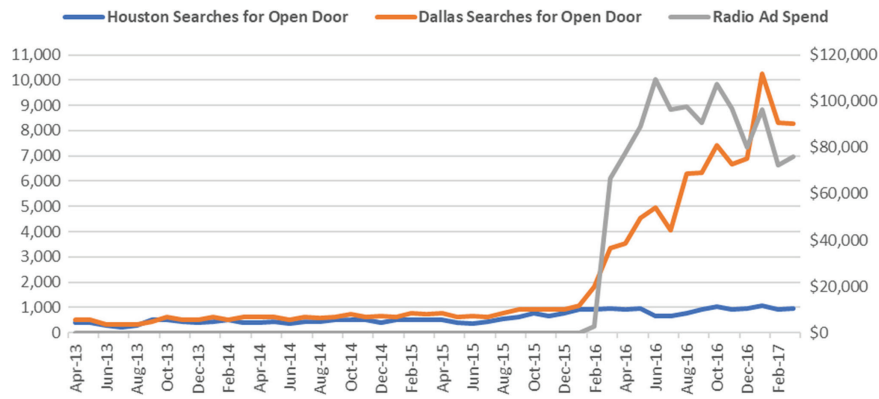

Figure 4.1: Open door searches and Ad spend.

purchases. Their results suggest that augmenting sales response models with online search data as a proxy for consumer purchase interest can lead to not only more accurate forecasts of product sales but also more accurate and diagnostic estimates of advertising effects.

4.2 An Application of Using Online Search Data for Market Response Modeling

In the rest of this section, we illustrate using online search data as response variables in one specific application. Figure 4.1 plots three monthly time series between April 2013 and March 2017: (a) the number of searches containing "open door" or "opendoor" in the Houston DMA, according to GAKP; (b) the number of searches containing "open door" or "opendoor" in the Dallas DMA; and (c) the amount of spend on radio ads in the Dallas DMA by Opendoor, an online company that buys and sells residential real estate (https://www.opendoor.com/).

Prior to February 2016, Opendoor did not run any ad campaigns in either Houston or Dallas. Neither did it buy or sell any residential properties in those two markets. In February 2016, Opendoor entered the Dallas market via a campaign that relied primarily on radio ads aimed at generating brand awareness among the target audience (e.g., low to middle-income households who are likely willing to sell their homes for cash) and having them search for the brand online and learn more

about how Opendoor's home buying process works. Between February 2016 and March 2017, Opendoor spent in total about 1.2 million dollars on radio ads in the Dallas DMA. It stayed out of the Houston market during this period.

How much did it cost Opendoor to generate one brand search via Google's search engine? Having an accurate estimate of the cost per brand search would allow Opendoor to figure out the breakeven ratio needed in converting brand searchers into customers who sell their homes to Opendoor for cash. To answer this question, we do the following attribution analysis.

First, we recognize that "open door" is a generic phrase that can indicate search interests unrelated to Opendoor the brand (e.g., open door policy). Second, Opendoor's ad campaigns in other markets that it had entered could have had spillover effects on the Dallas market. To account for these potential confounds, we can use "open door" and "opendoor" searches in the Houston DMA as a control to create a counterfactual for what the search volume could have been in the Dallas DMA if Opendoor had not entered the Dallas market with a radio ad campaign. To create the counterfactual, we regress searches in Dallas against searches in Houston, using data from April 2013 through January 2016, which leads to an intercept estimate of 113.84, a slope estimate of 1.0949, and an R-square of 0.776.

Equipped with the regression model for establishing the counterfactual, we fit the following ad response model to data between February 2016 and April 2017:

$$\text{Openddoor\_Search}_t^{\text{Dallas}} = 113.84 + 1.0949 \times \text{Openddoor\_Search}_t^{\text{Houston}} + \beta \times \text{Adstock}_t^{\text{Dallas}} + e_t \tag{4.1}$$

$$\text{Adstock}_t{}^{\text{Dallas}} = \lambda \times \text{Adstock}_{t-1}^{\text{Dallas}} + \text{Adspend}_t{}^{\text{Dallas}} \tag{4.2}$$

where e_t is assumed to be distributed i.i.d., normal with mean zero and standard deviation of σ_e; λ captures the monthly carryover rate of adstock; β captures the same-month effect of ad spend on Opendoor brand search; and $\frac{\beta}{1-\lambda}$ captures the long-term, cumulative impact of ad spend on Opendoor brand search.

We calibrate the model using nonlinear least squares. Figure 4.2 plots the model-predicted and the actual monthly Opendoor searches

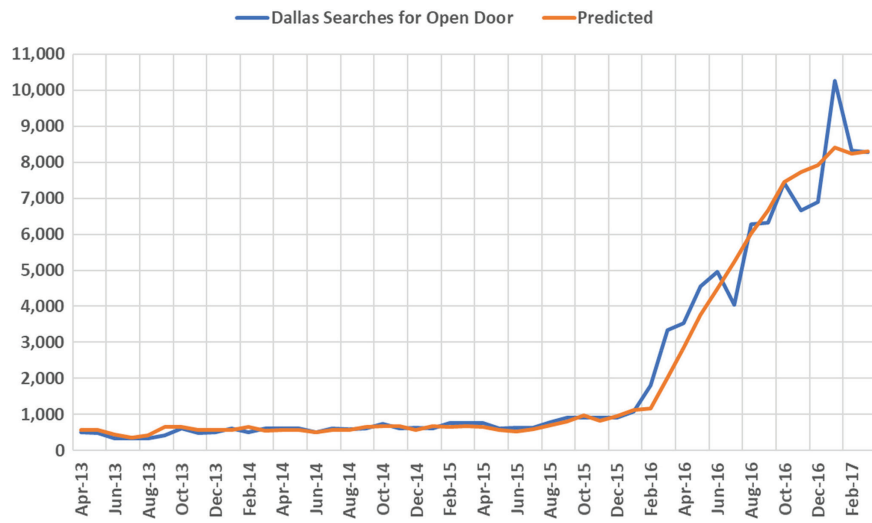

Figure 4.2: Open door searches: Actual vs. model-predicted.

in Dallas, which has an R-square of 0.847. λ is estimated to be 0.872 ($p < 0.01$), β is 0.01254 ($p < 0.01$), and $\frac{\beta}{1-\lambda}$ is 0.09786 ($p < 0.01$). In other words, for every \$1,000 spend on radio ads, Opendoor was able to generate on average about 98 ($=0.09786 * 1{,}000$) incremental brand searches, which translates to \$10.22 per incremental brand search. It is worth noting that the long-term impact on online brand search is 7.8 times as large as the same-month impact, indicating substantial carryover of offline media such as radio in building brand awareness and interest, as manifested in online brand searches.

In summary, the above application illustrates how online search data can be used as response variables to enrich our understanding of the impact of marketing on an integral stage of the modern customer journey—what they search online. Of course, in practice, researchers also need to address potential endogeneity threats (e.g., correlated unobservables) in order to establish the true causal impact of marketing on online searches (see Joo *et al.*, 2014, 2016; and Du *et al.*, 2019 for examples of identification strategies when online searches are treated as response variables).

5

Online Search Data as Proxies for Constructs of Interest

5.1 Review of Existing Studies

In this section, we review studies that do not fit neatly into either of the previous two streams of research. These studies use online search data creatively by treating what people search online as unvarnished reflections of the public psyche, uncovering what people really think, feel, and intend to do.

Compared to social media and survey data that can suffer from social desirability biases, online search data can be more reliable in capturing people's genuine attitudes and thoughts because search engines offer a platform for people to seek information with minimum self-censorship. Recognizing this, for example, a small stream of research has used aggregate online search data to measure levels of racism or sexism in different geographic areas, shedding light on sensitive topics in political science and public health where unvarnished data are hard to gather (Chae *et al.*, 2015, 2018; Connor *et al.*, 2019; Corbi and Picchetti, 2020; Stephens-Davidowitz, 2014).

Stephens-Davidowitz (2014) measures a region's racial animus by the proportion of Google searches containing racist phrases and examines

the effects of racial animus on voting behaviors. He finds that racial animus costed Obama roughly 4% of the national popular vote in both 2008 and 2012, which is 1.5 to 3 times larger than estimates based on survey measures of racial animus. Similarly, Chae *et al.* (2015, 2018) find positive associations between a region's racism and Black mortality rates and adverse birth outcomes such as preterm birth and low birth weight. These studies show that aggregate online search data could be used as alternative measures of sensitive constructs such as racial animus, which would help shed light on the antecedents and consequences of those variables.

Towers *et al.* (2015) examine the impact of mass media coverage of Ebola on online searches for Ebola-related information. They find that each Ebola-related news video can lead to tens of thousands of Ebola-related online searches. Alicino *et al.* (2015) suggest that Ebola-related online searches in most countries could have been attributed to unbalanced media coverage and the digital divide. Brodeur *et al.* (2021) estimate the impact of lockdowns during the COVID-19 pandemic on well-being-related online searches. They find a significant increase in searches for loneliness, worry, and sadness, along with a decrease in searches related to stress, suicide, and divorce during the lockdowns.

5.2 An Application of Using Online Search Data as Proxies for Hard-to-Measure Variables

The above studies illustrate how researchers may creatively use online search data as proxies for hard-to-measure variables. This GT example[1] shows how gifting behavior may vary depending on gender, marital status, and their interaction. In the rest of this section, we provide one specific application in more depth.

PepsiCo announced the removal of aspartame (an artificial sweetener) from its flagship Diet Pepsi in April 2015, citing results from large-scale surveys suggesting that the number one reason U.S. consumers

[1] https://trends.google.com/trends/explore?geo=US&q=gift%20for%20girlfriend%20%2B%20gifts%20for%20girlfriend,gift%20for%20boyfriend%20%2B%20gifts%20for%20boyfriend,gift%20for%20wife%2B%20gifts%20for%20wife,gift%20for%20husband%2B%20gifts%20for%20husband.

shunned diet colas was concerns about aspartame (Ester and Mickle, 2015). Unfortunately, that decision turned out to be a debacle akin to the 1985 "New Coke" fiasco—PepsiCo had to pull and replace the new aspartame-free Diet Pepsi with the original after three years of consumer backlash and declining sales (Kelso, 2018). The root cause of PepsiCo's blunder is that, despite the due diligence and market research accompanying such a high-stake decision, it mistook what consumers told market researchers in surveys (i.e., they were increasingly concerned about aspartame in diet colas) for what they truly thought and intended to do (i.e., they apparently were not concerned enough about aspartame to trade the original Diet Pepsi for an aspartame-free one).

Could PepsiCo have come to a different conclusion about aspartame being the driving force behind declining sales of diet colas by looking into online search data? It is conceivable that concerned consumers would seek information about aspartame online, and increasing concerns would manifest in increasing searches. To gather relevant search data, one needs first to identify search queries indicating concerns about aspartame. The query with the word "aspartame" by itself does not necessarily indicate that the searcher is concerned about aspartame. For example, a food science student could be looking for information about "aspartame" for a school project. To identify queries that are clearly concern-driven, we enter "aspartame" as the seed keyword into the "Discover new keywords" tool of GAKP, which produces a list of related keywords. We go through the list manually and select keywords that are deemed concern-driven. This process is repeated with newly identified keywords as seeds until no new relevant keywords are suggested by GAKP. The final list of keywords includes the word "aspartame" plus one of the following 28 words:

- Effects, poisoning, dangers, bad, cancer, without, safe, pregnancy, withdrawal, toxicity, risks, diabetes, detox, allergy, headaches, weight, pregnant, poison, symptoms, dangerous, safety, addiction, harmful, diarrhea, danger, insulin, warnings, free.

To pull historical search trend data for the above keywords, we use GT with the following two URLs: GT Query A[2] and GT Query B[3] (use the links to retrieve the raw GT data). Query A contains five composite keywords, and Query B contains two. They share the same first composite keyword, i.e., aspartame effects + aspartame poisoning + aspartame dangers + aspartame bad + aspartame cancer, which has the maximum GT index value of 100 among all the composite keywords in both Query A and Query B. As a result, we can add the index value of the second composite keyword in Query B (i.e., aspartame diarrhea + aspartame danger + aspartame insulin + aspartame warnings + aspartame free) to the sum of index values of the five composite keywords in Query A, resulting in the GT index value for our list of 28 keywords that indicate various concerns about aspartame. This workaround is necessary due to GT's cap on length per composite keyword and only up to five composite keywords can be simultaneously included in one GT query.

Figure 5.1 plots the monthly index values between January 2004 and June 2022. We see long-term trend, seasonality, and short-term fluctuations. To decompose the time series into those components so that we can see the long-term trend more clearly, we use the following unobserved components model:

[2] https://trends.google.com/trends/explore?date=all&geo=US&q=aspartame%20effects%20%2B%20aspartame%20poisoning%20%2B%20aspartame%20dangers%20%2B%20aspartame%20bad%20%2B%20aspartame%20cancer,aspartame%20without%20%2B%20aspartame%20safe%20%2B%20aspartame%20pregnancy%20%20%2B%20aspartame%20withdrawal,aspartame%20toxicity%20%2B%20aspartame%20risks%20%2B%20aspartame%20diabetes%20%2B%20aspartame%20detox%20%2B%20aspartame%20allergy,aspartame%20headaches%20%2B%20aspartame%20weight%20%2B%20aspartame%20pregnant%20%2B%20aspartame%20poison%20%2B%20aspartame%20symptoms,aspartame%20dangerous%20%2B%20aspartame%20safety%20%2B%20aspartame%20addiction%20%2B%20aspartame%20harmful.

[3] https://trends.google.com/trends/explore?date=all&geo=US&q=aspartame%20effects%20%2B%20aspartame%20poisoning%20%2B%20aspartame%20dangers%20%2B%20aspartame%20bad%20%2B%20aspartame%20cancer,aspartame%20diarrhea%20%2B%20aspartame%20danger%20%2B%20aspartame%20insulin%20%2B%20aspartame%20warnings%20%2B%20aspartame%20free.

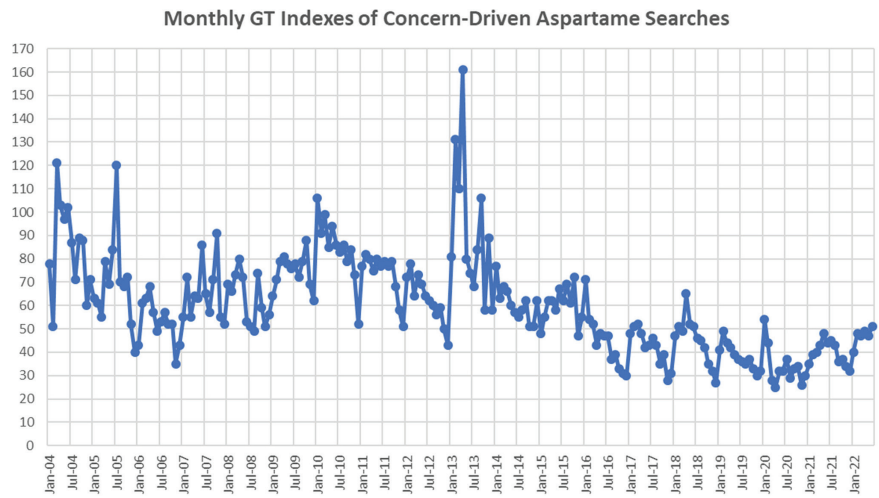

Figure 5.1: Concern-driven aspartame searches.

$$\ln(y_t) = \text{level}_t + \text{seasonality}_t + \text{autoregressive}_t + e_t \tag{5.1}$$

$$\text{level}_t = \text{level}_{t-1} + \varepsilon_t \tag{5.2}$$

$$\sum_{k=0}^{11} \text{seasonality}_{t-k} = \epsilon_t \tag{5.3}$$

$$\text{autoregressive}_t = \rho \times \text{autoregressive}_{t-1} + \omega_t \tag{5.4}$$

where y_t denotes the GT index value in month t, and level_t, seasonality_t, autoregressive_t, and e_t denote four latent components that represent, respectively, the level of the long-term trend in month t, seasonal deviation from the long-term trend in month t, a short-term autoregressive deviation from the long-term trend in month t, and white noise. The long-term trend is assumed to evolve following a random walk, with the standard deviation (σ_ε) of monthly level shifts to be empirically determined. The sum of seasonal deviations of twelve consecutive months is assumed to be stochastic, with a mean of zero and a standard deviation of σ_ϵ. The short-term autoregressive deviation is assumed to have a damping factor of ρ and a random monthly shock distributed i.i.d. normal with mean zero and standard deviation of σ_ω. The white noise is assumed to be distributed i.i.d. normal with mean zero and standard deviation of σ_e.

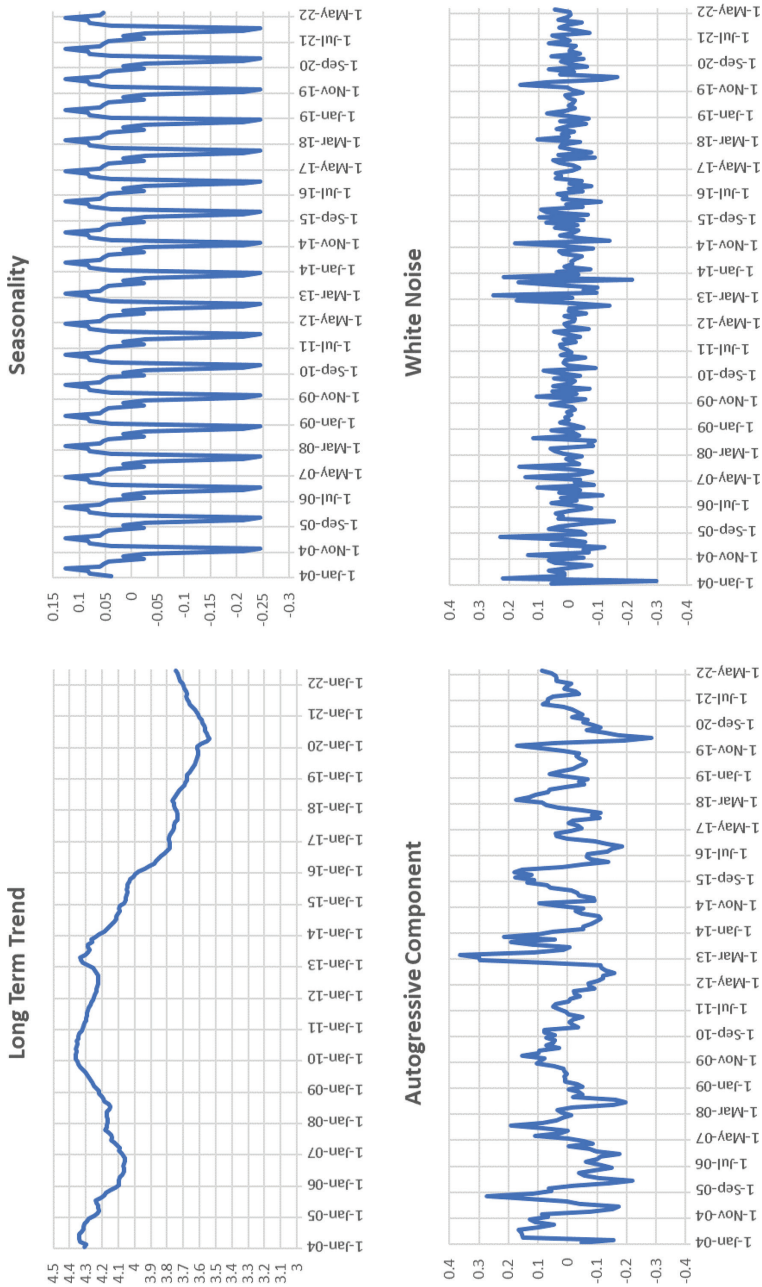

Figure 5.2: Decomposition of concern-driven aspartame searches.

We estimate the above model using the SAS procedure UCM. Figure 5.2 plots the model-inferred long-term trend, seasonality, short-term autoregressive, and white noise components. We see that the long-term trend peaked in January 2010 and had been in decline ever since until bottoming out in April 2020, a 56% decrease over a decade. In other words, based on concern-driven online searches of aspartame, we can say U.S. consumers became much less worried about aspartame in the five years prior to PepsiCo's 2015 decision to make Diet Pepsi aspartame free, which runs counter to what PepsiCo had concluded from their survey data. Online search data shows that the decline continued after 2015 for another five years.

In summary, through the above example, we illustrate how online search data may provide an unvarnished view of what consumers really think, feel and intend to do. The key lies in being creative and identifying search queries that are manifestations of the underlying topic of interest. To see the long-term trend more clearly, one needs to filter out seasonality and short-term fluctuations from the raw online search data, using a decomposition and smoothing model such as the one used in our example.

6

Ideas for Future Research

In the previous sections, we reviewed studies on how online search data can be used as predictors for nowcasting and forecasting, response variables for quantifying the impact of marketing, and proxies for capturing hard-to-measure consumer mindset and behavior. In this section, we identify several directions for more studies where we believe the potential of online search data has been under-tapped.

6.1 Using Online Search Data for Brand Health Tracking

Search queries containing brand names should be of particular interest to marketing researchers. Searching for a brand name by itself indicates a level of awareness, although one cannot tell whether the search is merely navigational (e.g., as a shortcut to get to the brand's website), informational, or transactional. Systematically examining what other terms are co-searched with the brand name can reveal more about the searcher's intent.

One type of co-searches can be quite informative of the dynamics of market structure. When consumers include two brand names in one query, e.g., "honda pilot vs toyota highlander," which is searched on average more than five thousand times per month by U.S. consumers,

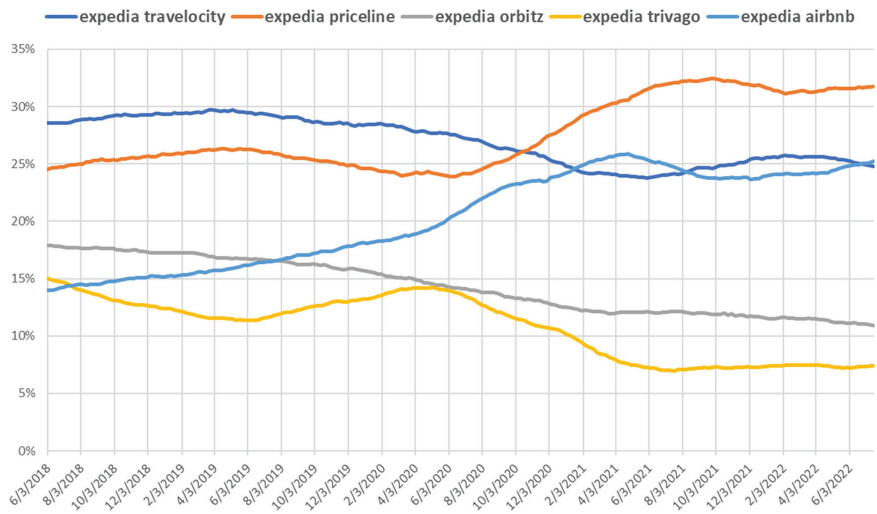

Figure 6.1: Trends in Expedia brand co-search shares.

they are likely conducting comparison shopping, indicating proximity of the two brands in the searcher's mind. By examining what other brands are most commonly co-searched with a focal brand, one can potentially identify the top direct competitors of the focal brand and quantify the relative pair-wise competitive intensity, which can be monitored over time and across markets.

In this GT example[1] (use the link to retrieve the raw data), we see how U.S. consumers have co-searched Expedia with five competitors in online travel shopping—Travelocity, Priceline, Orbitz, Trivago, and Airbnb—over a five-year period (June 2017 through July 2022). Figure 6.1 plots 52-week moving averages of the five competitors' shares among Expedia's brand co-searches, based on which we see that Expedia's top competitor in 2018 was Travelocity. By 2022, Priceline has replaced Travelocity as Expedia's top competitor. During the same time span, Airbnb has risen substantially as a competitive threat, going from the least co-searched to the second among the five Expedia competitors.

[1] https://trends.google.com/trends/explore?date=2017-06-04%202022-07-23&geo=US&q=expedia%20travelocity,expedia%20priceline,expedia%20orbitz,expedia%20trivago,expedia%20airbnb.

Taking the above example one step further, one can gather all the pair-wise co-search indexes, use the data as measures of brand similarity, and uncover the underlying market structure of the six competing brands, in a similar vein as Netzer *et al.* (2012), who use patterns of brand co-mentions in product reviews to uncover the market structure underneath competing brands.

In addition to being co-searched with competitors' brand names, brands are also co-searched with other terms such as product attributes (e.g., ford f150 mpg), price (e.g., 2022 f150 price), reviews, and locations. These co-searched terms provide clearer indications of the searcher's intent. By identifying terms that are commonly co-searched with a brand name, one can better understand consumers' brand information needs and brand associations.

To identify co-searched terms, one can use the focal brand name and its common variants as seed keywords in GAKP's "Discover new keywords" tool and screen for suggested keywords that include the focal brand name and some other terms. To organize these co-searched terms, which can number in hundreds or even thousands, one needs to come up with a tagging taxonomy with a smaller number of categories, which can be accomplished through either brute force or off-the-shelf text mining tools for automated topic discovery. Figure 6.2 shows the shares of product attributes co-searched with Chobani, a major brand of Greek yogurt. We see that fat is more frequently co-searched than protein, which is more frequently co-searched than sugar and carb. For the brand manager of Chobani, this information is potentially useful in product packaging and ad copy design. By monitoring over time and across markets the relative prevalence of product attributes co-searched with their brands, marketers can tailor their branding strategy more proactively to variation in brand associations.

In conclusion, online search data can serve as a big-data supplement in brand health tracking. A promising area of research would be to see how best to integrate patterns of online brand searches with survey and social media data in measuring and diagnosing brand health.

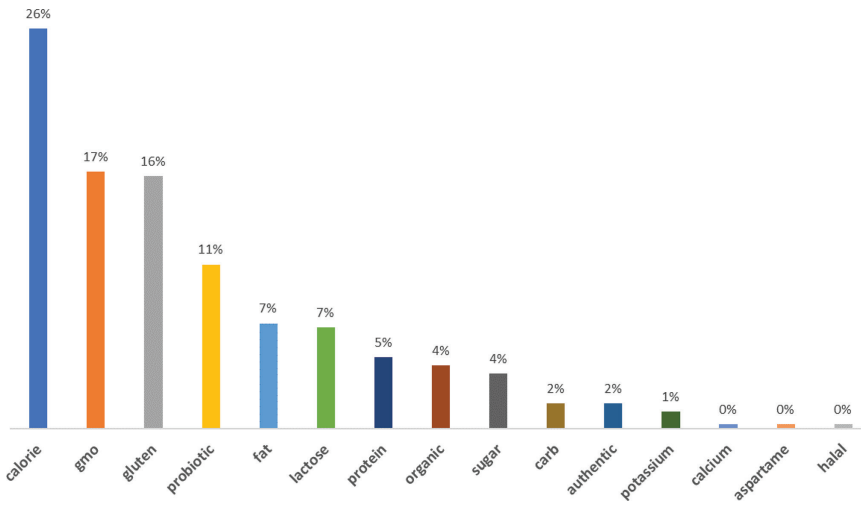

Figure 6.2: Shares of product attributes co-searched with Greek Yogurt Brand Chobani.

6.2 Using Online Search Data for Trendspotting

A key opportunity for growth lies in spotting emerging trends in market demands before the competition and adjusting marketing strategies of existing offerings or developing new offerings to leverage those trends (Du *et al.*, 2021). Because changing customer needs and wants can manifest in shifts in online searches (Du *et al.*, 2015), online search data, systematically gathered and analyzed, can play a critical role in trendspotting that informs growth opportunities (Du and Kamakura, 2012).

The biggest challenge in trendspotting is that, without the benefit of hindsight, it is difficult to distinguish short-lived fads from emergent trends that offer meaningful growth opportunities over the long run. When companies need to invest significant resources to develop and market a new product, mistaking a fad for a trend could prove acutely detrimental. This challenge is particularly salient in industries such as beauty, fashion, and food, where consumer preferences and behaviors

are constantly changing (Dubois and Bens, 2014; Google, 2019). Consequently, it is risky to implement a proactive growth strategy contingent on identifying emerging trends in consumer needs and wants.

Take gluten-free and ketogenic diets as two contrasting examples. Gluten-free foods have grown from a niche category into a multi-billion dollar business where avoiding gluten is now a lifestyle choice (Mintel Group, 2018), despite the lack of scientific evidence for its benefits to the general public (Levinovitz, 2015). Gluten-free products have become a source of sustained growth for many consumer-packaged goods manufacturers (Packaged Facts, 2016). In contrast, the ketogenic diet (or keto diet for short) is a high-fat, adequate-protein, low-carbohydrate dietary therapy that in medicine is used mainly to treat hard-to-control epilepsy children. However, many consumers have adopted keto diets for weight loss, despite the fact that whether it works in the long term or it is safe has not been established scientifically. Online searches[2] (use the link to retrieve the raw GT data) show that gluten free is a long-term trend that first emerged over a decade ago and has since experienced sustained growth and is here to stay. On the other hand, online searches for keto diets rose sharply between 2017 and 2018 and have since been in decline. By mid-2022, online searches for keto diets returned to the level in 2016, prior to its meteoric takeoff, suggesting that consumer interest in keto diets is in all likelihood a short-lived fad.

The above contrast between a long-term growth trend and a short-lived fad points to an important research question: How can marketers have the foresight to tell them apart as early as possible? Admittedly, projecting the trajectory of a trend or fad is a fundamental challenge in time series forecasting, especially during the pre-takeoff period when data is limited and accurate long-range forecasts are highly valuable (Chandrasekaran and Tellis, 2007).

One way to solve this "cold start" problem is to build a large training sample of historical trends and fads that cover a wide range of industries, pairing it with pattern recognition methods to identify the most similar historical counterparts to help make forecasts for the

[2] https://trends.google.com/trends/explore?date=all&geo=US&q=gluten%20free,keto%20diet%20%2B%20keto%20diets%20%2B%20ketogenic%20diet%20%2B%20ketogenic%20diets.

emergent trend or fad (Heist and Tarraf, 2016). Unfortunately, such training samples are hard to come by, the construction of which will require painstaking efforts. Online search data offers an easy-to-access, comprehensive source for doing so by providing long historical data that quantifies the magnitude and momentum of trends and fads at any given moment in time, including the pre-takeoff period. One may follow the principles that have been successfully applied to discover empirical regularities in the diffusion of technological innovations and new products (Golder *et al.*, 2009). In combination with big training samples, new econometric or machine learning methods for conducting large-scale trend analysis are needed to better separate emergent trends from fads as early as possible and to predict their long-run trajectories as accurately as possible.

One key feature of online search data is that it is available by geographic areas such as cities, states, and countries. Because the timing with which a trend or fad unfolds may vary across markets, some of which could be "harbingers" that send early-warning signals about what is to come (Anderson *et al.*, 2015). Identifying these trendsetting markets could help spot the rise or fall of a trend or fad sooner.

For example, relative to granite countertops, the popularity of quartz countertops in the U.S. has grown steadily and substantially over the last decade, as manifested in GT data[3] (use the link to retrieve the raw data). Figure 6.3 plots the twelve-month moving average of quartz countertop search to granite countertop search ratio from December 2014 through June 2022. We see that the ratio hovered around 15% until January 2011 and went on continued growth ever since. By 2022, the ratio reached 90%, six times that of 2011. Furthermore, the growth of quartz countertop popularity is by no means uniform across markets. Figure 6.4 plots the ratios across different states as of June 2022. We see that quartz countertops are far more popular in the west coast states (e.g., with a ratio of 170% in California and Washington), North Dakota (186%), and the rest of the Midwest. In contrast, they remain much less searched than granite countertops in both southwest (e.g., a ratio

[3] https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fg%2F11h79tqt8_,%2Fg%2F11fl7dx7t6.

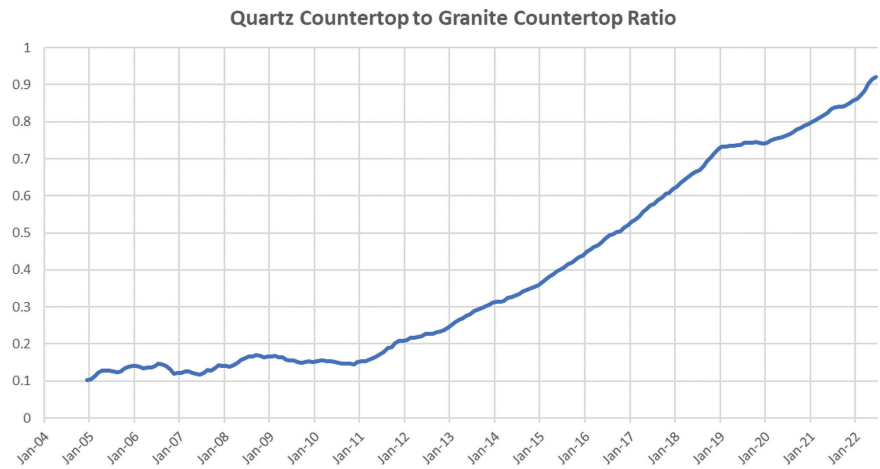

Figure 6.3: Growth of online searches for quartz countertop relative to granite countertop.

of 61% in Louisiana) and southeast (e.g., a ratio of 67% in Mississippi and Alabama). Such a large variation in the relative popularity of quartz countertops across markets suggests that the growth trend of quartz countertops has been led by the west coast and midwest states. Therefore, by monitoring online search trends for quartz countertops in the west coast and midwest states, one could potentially foresee what is to come in the southwest and southeast states.

Finally, cutting across national boundaries, online search data offers rich opportunities for studying how trends and fads propagate across countries, which can prove particularly useful in the context of international marketing. For instance, trends may spread via both online and offline word of mouth, with offline contagion relying more on geographic proximity, while online contagion more easily travels around the world, as manifested in geo-temporal correlation in online search patterns[4] (click the link to see how the evolution of online searches for quartz countertops has differed across the U.S. Canada, U.K., and Australia). A diffusion model that distinguishes between online and offline contagion

[4] https://trends.google.com/trends/explore?date=all,all,all,all&geo=US,CA,GB,AU&q=%2Fg%2F11fl7dx7t6,%2Fg%2F11fl7dx7t6,%2Fg%2F11fl7dx7t6,%2Fg%2F11fl7dx7t6.

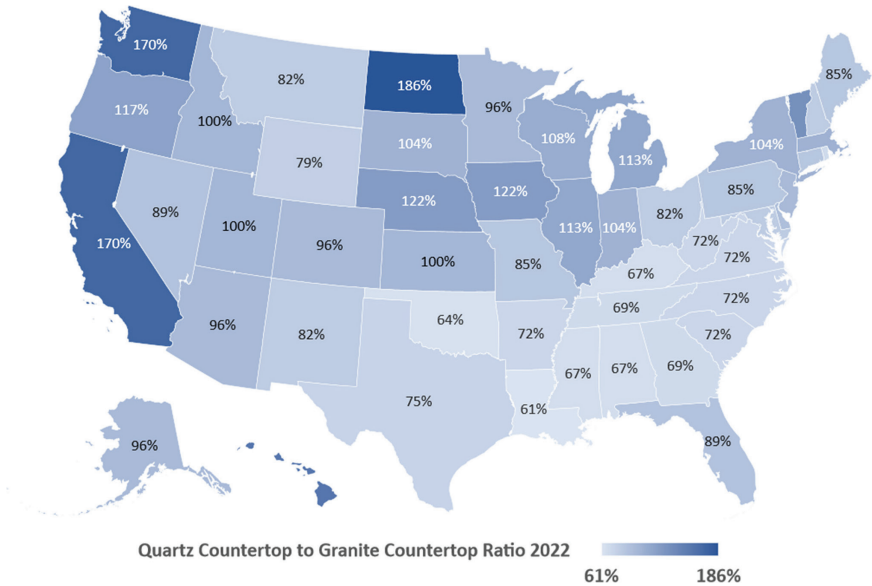

Figure 6.4: Cross-market differences in online searches for quartz countertop relative to granite countertop.

may help identify how quickly a trend spreads both across and within countries.

6.3 Using Online Search Data in Behavioral Research

For behavioral research, online search data can potentially help identify new phenomena for investigation and augment findings from experimental data. For example, Kozinets *et al.* (2017) utilize Google Trends data to illustrate rising consumer interest in "food porn" (images of unhealthy foods). Ross *et al.* (2020) use Google Trends data to motivate their research by showing the growth of consumer searches for keywords related to "downsizing" and "decluttering." Galoni *et al.* (2020) use data from Google Flu Trends to show empirical evidence on how the presence of contagious disease influences what and how consumers buy, corroborating their findings from experimental studies.

Online search data, when utilized creatively, can also shed light on behavioral constructs that may prove hard to quantify through surveys

or manipulate in experimental settings. Stephens-Davidowitz (2017) provides many intriguing examples in this regard (e.g., racial animus, gender animus, depression, sex and sexual orientation, Islamophobia, self-induced abortion, drug abuse).

6.4 Limitations of Online Search Data

One major threat to online search data as a source of marketing insights lies in the fact that how consumers search for information constantly evolves with technological advances. Traditional search engines such as Google and Baidu are no longer the default option for many information seekers nowadays. Increasingly specialized platforms and websites provide their own search functions that offer consumers with more relevant information in a specific product category or domain. For example, consumers can acquire information about fashionable topics directly from social media (e.g., https://trends.pinterest.com/) or learn about restaurants or tourist locations from review platforms such as Yelp and TripAdvisor. Consequently, the online search market may become more fragmented, making data from any single platform less representative of the needs, wants and wishes of the general population.

More recently, the introduction and popularization of large language models such as ChatGPT, which can better recognize natural language queries and provide summarized responses, has the potential of completely reshaping how consumers seek information on the Internet. One implication would be that researchers can no longer rely on keywords as proxies of intents; rather, more sophisticated natural language processing methods will be needed in order to decipher unstructured queries made by consumers. To us, this presents both a threat and an opportunity for learning about consumers from what and how they search online.

7

Concluding Remarks

The main objective of this monograph is to make the case that online search data from sources such as GT and GAKP can and should be leveraged by both marketing academics and practitioners for marketing insights and foresights. We start by offering a brief tutorial of Google Trends and Google Ads Keyword Planner, focusing on lesser-known features and offering tips that we have found particularly useful in practice in order to get the most out of these platforms.

Our review of the literature follows three threads. First, we survey research that has treated aggregate online search interests as either concurrent or leading indicators of real-world phenomena. This stream of research focuses mainly on gauging the value of online search data as predictors in improving the performance of either nowcasting or forecasting. Second, we examine research that has treated aggregate online searches as response variables that can help measure and improve marketing effectiveness in terms of both immediate and longer-term impacts. Third, we review research that has treated patterns of online searches as unvarnished reflections of the public psyche, uncovering what people really think, feel, and intend to do, insights that may otherwise

be difficult to ascertain based on what people post on social media or tell market researchers in surveys.

We highlight a couple of areas for future research where online search data can serve as a big-data supplement to traditional market research: brand health tracking and trendspotting. We argue that competitor brands and product attributes co-searched with a focal brand can tell us a lot about the competitive landscape and brand association and how those key elements of brand health vary across markets and evolve over time. We believe that online search data, properly mined, can help marketers spot emergent trends in consumer needs and wants that can reshape market boundaries while separating them from fleeting fads.

To conclude, we hope more marketing researchers will leverage online search data, in increasingly rigorous and creative ways, as an integral part of modern marketing insights systems, just like how they have embraced social media data, click stream data, CRM data, scanner panel data, single-source data, syndicated survey data, etc.

References

Alicino, C., N. L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, and A. Orsi (2015). "Assessing Ebola-related web search behaviour: Insights and implications from an analytical study of Google Trends-based query volumes". *Infectious Diseases of Poverty*. 4: 54. DOI: 10.1186/s40249-015-0090-9.

Anderson, E., S. Lin, D. Simester, and C. Tucker (2015). "Harbingers of failure". *Journal of Marketing Research*. 52(5): 580–592. DOI: 10.1509/jmr.13.0415.

Askitas, N. and K. Zimmermann (2009). "Google econometrics and unemployment forecasting". *Applied Economics Quarterly*. 55: 107–120. DOI: 10.3790/aeq.55.2.107.

Bangwayo-Skeete, P. F. and R. W. Skeete (2015). "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach". *Tourism Management*. 46: 454–464. DOI: 10.1016/j.tourman.2014.07.014.

Bijl, L., G. Kringhaug, P. Molnár, and E. Sandvik (2016). "Google searches and stock returns". *International Review of Financial Analysis*. 45: 150–156. DOI: 10.1016/j.irfa.2016.03.015.

Brodeur, A., A. E. Clark, S. Fleche, and N. Powdthavee (2021). "COVID-19, lockdowns and well-being: Evidence from Google Trends". *Journal of Public Economics*. 193: 104346. DOI: 10.1016/j.jpubeco.2020.104346.

Brynjolfsson, E., T. Geva, and S. Reichman (2016). "Crowd-Squared: Amplifying the predictive power of search trend data". *MIS Quarterly*. 40(4): 941–962. DOI: 10.25300/MISQ/2016/40.4.07.

Chae, D., S. Clouston, M. L. Hatzenbuehler, M. R. Kramer, H. L. Cooper, S. M. Wilson, S. Stephens-Davidowitz, R. S. Gold, and B. G. Link (2015). "Association between an internet-based measure of area racism and black mortality". *PLoS One*. 10(4): e0122963. DOI: 10.1371/journal.pone.0122963.

Chae, D. H., S. Clouston, C. D. Martz, M. L. Hatzenbuehler, H. Cooper, R. Turpin, S. Stephens-Davidowitz, and M. R. Kramer (2018). "Area racism and birth outcomes among blacks in the United States". *Social Science & Medicine*. 199: 49–55. DOI: 10.1016/j.socscimed.2017.04.019.

Chan, E. H., V. Sahai, C. Conrad, and J. S. Brownstein (2011). "Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance". *PLoS Neglected Tropical Diseases*. 5(5): 1206. DOI: 10.1371/journal.pntd.0001206.

Chandrasekaran, D., R. Srinivasan, and D. Sihi (2018). "Effects of offline ad content on online brand search: Insights from super bowl advertising". *Journal of the Academy of Marketing Science*. 46(3): 403–430. DOI: 10.1007/s11747-017-0551-8.

Chandrasekaran, D. and G. J. Tellis (2007). "A critical review of marketing research on diffusion of new products". *Review of Marketing Research*: 39–80.

Choi, H. and H. Varian (2009). *Predicting initial claims for unemployment benefits*. Research report, Google Inc. URL: http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/initialclaimsUS.pdf.

Choi, H. and H. Varian (2012). "Predicting the present with Google Trends". *Economic Record*. 88(1): 2–9. DOI: 10.1111/j.1475-4932.2012.00809.x.

Connor, P., V. Sarafidis, M. J. Zyphur, D. Keltner, and S. Chen (2019). "Income inequality and white-on-black racial bias in the United States: Evidence from project implicit and Google Trends". *Psychological Science*. 30(2): 205–222. DOI: 10.1177/0956797618815441.

Corbi, R. and P. Picchetti (2020). "The cost of gendered attitudes on a female candidate: Evidence from Google Trends". *Economics Letters*. 196: e109495. DOI: 10.1016/j.econlet.2020.109495.

Da, Z., J. Engelberg, and P. Gao (2011). "In search of attention". *The Journal of Finance*. 66(5): 1461–1499. DOI: 10.1111/j.1540-6261.2011.01679.x.

D'Amuri, F. and J. Marcucci (2017). "The predictive power of Google searches in forecasting US unemployment". *International Journal of Forecasting*. 33(4): 801–816. DOI: 10.1016/j.ijforecast.2017.03.004.

Dimpfl, T. and S. Jank (2016). "Can internet search queries help to predict stock market volatility?" *European Financial Management*. 22(2): 171–192. DOI: 10.1111/eufm.12058.

Du, R. Y., Y. Hu, and S. Damangir (2015). "Leveraging trends in online searches for product features in market response modeling". *Journal of Marketing*. 79(1): 29–43. DOI: 10.1509/jm.12.0459.

Du, R. Y. and W. A. Kamakura (2012). "Quantitative trendspotting". *Journal of Marketing Research*. 49(4): 514–536. DOI: 10.1509/jmr.10.0167.

Du, R. Y., O. Netzer, D. Schweidel, and D. Mitra (2021). "Capturing marketing information to fuel growth". *Journal of Marketing*. 85(1): 163–183. DOI: 10.1177/0022242920969198.

Du, R. Y., L. Xu, and K. C. Wilbur (2019). "Immediate responses of online brand search and price search to TV ads". *Journal of Marketing*. 83(4): 81–100. DOI: 10.1177/0022242919847192.

Dubois, D. and K. Bens (2014). "Ombre, tie-dye, splat hair: Trends or fads? 'Pull' and 'Push' social media strategies at L'Oréal Paris". *Instead*.

Dugas, A. F., Y. H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman (2012). "Google flu trends: Correlation with emergency department influenza rates and crowding metrics". *Clinical Infectious Diseases*. 54(4): 463–469. DOI: 10.1093/cid/cir883.

Ester, M. and T. Mickle (2015). "Pepsico to drop aspartame from diet pepsi: Consumer backlash, slumping sales prompt beverage giant to switch artificial sweeteners". *Wall Street Journal*. April 24.

Ettredge, M., J. Gerdes, and G. Karuga (2005). "Using web-based search data to predict macroeconomic statistics". *Communications of the ACM*. 48: 87–92. DOI: 10.1145/1096000.1096010.

Galoni, C., G. S. Carpenter, and H. Rao (2020). "Disgusted and afraid: Consumer choices under the threat of contagious disease". *Journal of Consumer Research*. 47(3): 373–392. DOI: 10.1093/jcr/ucaa025.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2009). "Detecting influenza epidemics using search engine query data". *Nature*. 457(7232): 1012–1014. DOI: 10.1038/nature07634.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010). "Predicting consumer behavior with web search". *Proceedings of the National Academy of Sciences*. 107(41): 17486–17490. DOI: 10.1073/pnas.1005962107.

Golder, P. N., R. Shacham, and D. Mitra (2009). "Findings—Innovations' origins: When, by whom, and how are radical innovations developed?" *Marketing Science*. 28(1): 166–179. DOI: 10.1287/mksc.1080.0384.

Google (2019). "What fashion fans around the world are searching for on Google". URL: https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/google-fashion-search-data/.

GS Statcounter (2022). "Search engine market share worldwide". On July 29, 2022. URL: https://gs.statcounter.com/search-engine-market-share.

Guitart, I. A. and S. Stremersch (2020). "The impact of informational and emotional television ad content on online search and sales". *Journal of Marketing Research*. 58(2): 299–320. DOI: 10.1177/0022243720962505.

Harvey, A. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge New York: Cambridge University Press.

Heist, G. and S. Tarraf (2016). "Trend analytics: A data-driven path to foresight". *Marketing Insights*: 18–19. Spring.

Hu, Y., R. Y. Du, and S. Damangir (2014). "Decomposing the impact of advertising: Augmenting sales with online search data". *Journal of Marketing Research*. 51(3): 300–319. DOI: 10.1509/jmr.12.0215.

Internet Live Stats (2022). "Google searches in 1 second". URL: https://internetlivestats.com/one-second/#google-band.

Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu (2014). "Television advertising and online search". *Management Science*. 60(1): 56–73. DOI: 10.1287/mnsc.2013.1741.

Joo, M., K. C. Wilbur, and Y. Zhu (2016). "Effects of TV advertising on keyword search". *International Journal of Research in Marketing*. 33(3): 508–523. DOI: 10.1016/j.ijresmar.2014.12.005.

Kelso, A. (2018). "Diet Pepsi is bringing Aspartame back, again". URL: https://www.fooddive.com/News/Diet-Pepsi-Is-Bringing-Aspartame-Back-Again/517618/.

Kozinets, R., A. Patterson, and R. Ashman (2017). "Networks of desire: How technology increases our passion to consume". *Journal of Consumer Research*. 43(5): 659–682. DOI: 10.1093/jcr/ucw061.

Kristoufek, L. (2013). "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era". *Scientific Reports*. 3(1): 3415. DOI: 10.1038/srep03415.

Kulkarni, G., P. K. Kannan, and W. Moe (2012). "Using online search data to forecast new product sales". *Decision Support Systems*. 52(3): 604–611. DOI: 10.1016/j.dss.2011.10.017.

Laroche, M., I. Kiani, N. Economakis, and M. O. Richard (2013). "Effects of multi-channel marketing on consumers' online search behavior". *Journal of Advertising Research*. 53(4): 431. DOI: 10.2501/JAR-53-4-431-443.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). "The parable of Google Flu: Traps in big data analysis". *Science*. 343(6176): 1203–1205. DOI: 10.1126/science.1248506.

Levinovitz, A. (2015). *The Gluten Lie: And Other Myths About What You Eat*. Regan Arts.

Li, X., B. Pan, R. Law, and X. Huang (2017). "Forecasting tourism demand with composite search index". *Tourism Management*. 59: 57–66. DOI: 10.1016/j.tourman.2016.07.005.

Mintel Group (2018). *US Gluten-Free Foods Market Report*. Available for purchase from https://store.mintel.com/.

Moz (2022). "We surveyed 1,400 searchers about Google—Here's what we learned". URL: https://moz.com/blog/new-google-survey-results.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012). "Mine your own business: Market-structure surveillance through text mining". *Marketing Science*. 31(3): 521–543. DOI: 10.1287/mksc.1120.0713.

Packaged Facts (2016). *Industry Report on Gluten-Free Foods in the U.S.* 6th edition. Available for purchase from https://www.freedoniagroup.com/packaged-facts/gluten-free-foods-in-the-us.

Pelat, C., T. Clément, A. Bar-Hen, A. Flahault, and A. J. Valleron (2009). "More diseases tracked by using Google Trends". *Emerging Infectious Diseases*. 15(8): 1327–1328. DOI: 10.3201/eid1508.090299.

Polgreen, P. M., Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein (2008). "Using internet searches for influenza surveillance". *Clinical Infectious Diseases*. 47(11): 1443–1448. DOI: 10.1086/593098.

Preis, T., H. S. Moat, and H. E. Stanley (2013). "Quantifying trading behavior in financial markets using Google Trends". *Scientific Reports*. 3(1): 1684. DOI: 10.1038/srep01684.

Preis, T., D. Reith, and H. E. Stanley (2010). "Complex dynamics of our economic life on different scales: Insights from search engine query data". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 368(1933): 5707–5719. DOI: 10.1098/rsta.2010.0284.

Reiley, D. and R. Lewis (2013). "Down-to-the-minute effects of super bowl advertising on online search behavior". In: *Proceedings of the ACM Conference on Electronic Commerce*.

Ross, G., M. G. Meloy, and L. E. Bolton (2020). "Disorder and downsizing". *Journal of Consumer Research*. 47(6): 959–977. DOI: 10.1093/jcr/ucaa051.

Santillana, M., A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein (2015). "Combining search, social media, and traditional data sources to improve influenza surveillance". *PLOS Computational Biology*. 11(10): e1004513. DOI: 10.1371/journal.pcbi.1004513.

Seifter, A., A. Schwarzwalder, K. Geis, and J. Aucott (2010). "The utility of Google Trends for epidemiological research: Lyme disease as an example". *Geospatial Health*. 4(2): 135–137. DOI: 10.4081/gh.2010.195.

Stephens-Davidowitz, S. (2014). "The cost of racial animus on a black candidate: Evidence using Google search data". *Journal of Public Economics*. 118: 26–40. DOI: 10.1016/j.jpubeco.2014.04.010.

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York NY: HarperCollins Publishers.

Teng, Y., D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong (2017). "Dynamic forecasting of Zika epidemics using Google Trends". *PLOS ONE*. 12(1): e0165085. DOI: 10.1371/journal.pone.0165085.

Towers, S., S. Afzal, G. Bernal, N. Bliss, S. Brown, B. Espinoza, J. Jackson, J. Judson-Garcia, M. Khan, M. Lin, R. Mamada, V. M. Moreno, F. Nazari, and C. Castillo-Chavez (2015). "Mass media and the contagion of fear: The case of Ebola in America". *PLOS ONE*. 10(6): e0129179. DOI: 10.1371/journal.pone.0129179.

Vaughan, L. and Y. Chen (2014). "Data mining from web search queries: A comparison of Google Trends and baidu index". *Journal of the Association for Information Science and Technology*. 66: 13–22. DOI: 10.1002/asi.23201.

Vosen, S. and T. Schmidt (2011). "Forecasting private consumption: Survey-based indicators vs. Google Trends". *Journal of Forecasting*. 30(6): 565–578. DOI: 10.1002/for.1213.

Wu, L. and E. Brynjolfsson (2009). "The future of prediction: How Google searches foreshadow housing prices and quantities". In: *ICIS, 2009 Proceedings 147*.

Xiong, G. and S. Bharadwaj (2014). "Prerelease buzz evolution patterns and new product performance". *Marketing Science*. 33(3): 401–421. DOI: 10.1287/mksc.2013.0828.

Yang, S., M. Santillana, and S. C. Kou (2015a). "Accurate estimation of influenza epidemics using Google search data via ARGO". *Proceedings of the National Academy of Sciences*. 112(47): 14473–14478. DOI: 10.1073/pnas.1515373112.

Yang, X., B. Pan, J. A. Evans, and B. Lv (2015b). "Forecasting Chinese tourist volume with search engine data". *Tourism Management*. 46: 386–397. DOI: 10.1016/j.tourman.2014.07.019.

Yu, L., Y. Zhao, L. Tang, and Z. Yang (2019). "Online big data-driven oil consumption forecasting with Google Trends". *International Journal of Forecasting*. 35(1): 213–223. DOI: 10.1016/j.ijforecast.2017.11.005.

Zigmond, D. and H. Stipp (2010). "Assessing a new advertising effect". *Journal of Advertising Research*. 50(2): 162. DOI: 10.2501/S0021849910091324.