

# Data Voids: Where Missing Data Can Easily Be Exploited

---

May 2018

MICHAEL GOLEBIEWSKI  
Microsoft Bing

DANAH BOYD  
Microsoft Research and Data & Society

## Contents

Introduction .....	1
How Search Engines Work .....	1
Data Voids as Vulnerabilities .....	2
Typology of Data Voids .....	6

## Introduction

Search plays a unique role in modern online information systems. Unlike people’s use of social media, where they primarily consume algorithmically curated feeds of information, people’s approaches to search engines typically begin with a query or question in an effort to seek new information. However, not all search queries are equal. Many more people search for “basketball” than “underwater basket weaving.” Likewise, a lot more content is created about the sport than the absurdist activity (although the latter’s pictures are pretty great!) As a result, when search engines like Bing and Google try to provide users with information about basketball, they have *a lot* more data to work with.

There are many search terms for which the available relevant data is limited, non-existent, or deeply problematic. We call these “data voids.” Most of these searches are rare, but in the cases where people do search for these terms, search engines tend to return results that may not give the user what they want because of limited data and/or limited lessons learned through previous searches. If you type a random set of characters into a search engine – e.g., “aslkfjastowerk;asndf” – you will probably return no results—simply because no pages contain that random set of letters. But there is a long tail between a term like “basketball,” which promises a seemingly infinite number of results, and one with zero results. In that long tail, there are plenty of search queries that can drop people into a data void rife with problematic results.

In this paper, we want to offer some basic background on search engines before discussing the different types of data voids; the challenges that search engines face when they encounter queries over spaces where data voids exist; and the ways data voids can be exploited by those with ideological, economic, or political agendas.

## How Search Engines Work

In order to organize information and respond to queries in a reliable manner, search engines like Bing and Google draw on available information (URLs and their content, links, images, videos, etc.) to build models that allow them to quickly identify and prioritize content that most likely matches the desired goals of a searcher. This isn’t an easy task, in no small part because what people search for is often vague. Indeed, machine learning technologies don’t ever understand intention; they simply build statistical models based on previous patterns found in training data that are used to define a successful outcome.

What is a user really trying to get at with a query like “subway?” Information about the closest transit station and its closures? Information about the fast food restaurant and its hours? A history of subways around the world?

The architects of search engines draw on any available information to help maximize the likelihood that search results give users what they want. The system underpinning search engines is designed to try and match a user’s intentions with results by drawing on many signals. These signals come from sources including the pages themselves (e.g., text on the page, anchor text, title of the page), searches and interactions from other users, and additional information like the geographical location of the person’s computer for queries like restaurants. Search interfaces are designed to try and coax the user into offering a bit more information by suggesting additional phrases in the auto-suggest bar, attempting to narrow the query.

Search engines aren’t human. They are machine learning systems designed by people; those systems have a limited ability to understand the *meaning* of words. They focus on the probabilistic likelihood that a given page, image, video, or news story will most likely yield a positive interaction for a user given their search query. Overlaps like the one between Subway-the-sandwich-shop and subway-the-underground-train aren’t identified through manual demarcation, but through statistical probabilities and models derived from the data that suggests a different topological link structure and related word context. For example, very few pages discussing the timetables of the train have detailed descriptions of sandwiches and their toppings.

Ideally, using a search engine will be easy for someone seeking information; perhaps the user won’t even need to click again to get what they want. For example, a search for basketball might offer up the NBA schedule and recent scores even before it shows links to webpages. It might provide a call-out from Wikipedia to offer broader overview information. In addition to the natural search results, the first page might include videos, images, news results, advertisements, and related searches, all intended to help users get to what they might possibly want.

On Bing, for instance, the overwhelming majority of users only engage with the first page of results. Less than 3% of queries result in someone going to the next page. Thus, what matters most is what is prioritized on the first page.

## Data Voids as Vulnerabilities

When search engines have little natural content to return for a particular query, they are more likely to return low quality and problematic content. This is because there is little high-quality content for the search engine to return.

Consider searching for “Harrold-Oklaunion.” Both Bing and Google promise you tens of thousands of results, but the front page of each is pretty barren. Most of it is algorithmically generated content by services like Accuweather.com, City-Data.com, Yellowpages.com, and Acrevalue.com. These are services that produce a unique page for every town in the country. There’s a smattering of links to Wikipedia entries, news stories, and court records. But all told, the results associated with this small Texas town of roughly 500 people are limited. Most likely, few people search for this town. Over the month of February 2018, there were so few queries run on Bing for this search term we could only find a very small number, likely from us. Given both limited data and limited searches, it would not be hard to radically alter the results. By creating new content with the town’s name and engaging in a small amount of search engine optimization, someone could with relative ease get search engines to return their website on the front page of results.

Some users come across data voids naturally while searching, but, in many cases, they are *guided* to these problematic search terms. Those seeking to manipulate information often prepare to use these terms by creating websites and leveraging basic search engine optimization techniques. They then encourage people to search for these terms by leaving trails on other pages or talking about a phrase in other media. For example, after creating a splashy website about Harrold-Oklaunion, someone with an agenda could encourage potential users to search for that town’s name through talk radio or by altering Wikipedia entries to indicate that someone is from there.

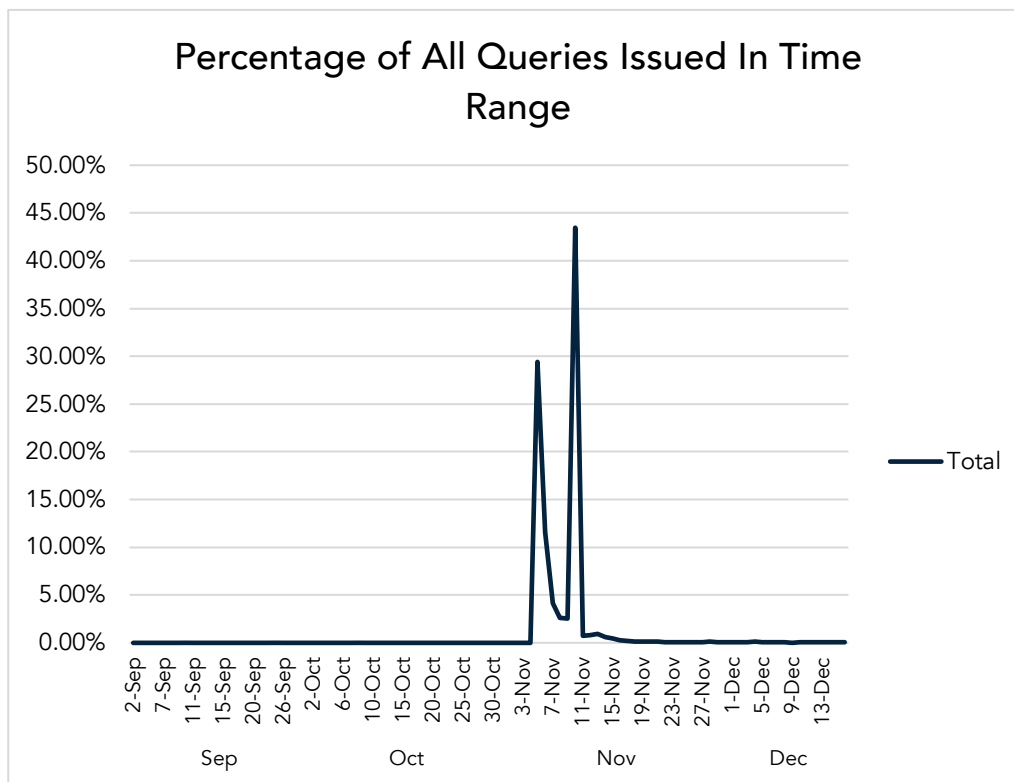
A rising trend in online misinformation is to encourage users to search for a topic for which the motivated manipulator knows that only one point of view will be represented. An example of this is where terms are strategically placed on websites to serve as a “redpill” in order to encourage curious web surfers to be exposed to hateful ideas. Nowhere is this more recognizable than through the story of Dylann Roof, who on June 17, 2015, after attending a bible study at the Emanuel African Methodist Episcopal Church in Charleston, S.C., opened fire and killed nine African American church-goers and injured one more. His act of terrorism and hatred is especially profound, given that he left a manifesto that described his radicalization towards white supremacy:

The event that truly awakened me was the Trayvon Martin case. I read the Wikipedia article and right away I was unable to understand what the big deal was. It was obvious that Zimmerman was in the right. But more importantly this prompted me to type in the words “black on white crime” into Google, and I have never been the same since that day. The first website I came to was the Council of Conservative Citizens. There were pages upon pages of these brutal black on white murders. I was in disbelief. At this moment I realized that

something was very wrong. How could the news be blowing up the Trayvon Martin case while hundreds of these black on white murders got ignored?

The term “black on white crimes” is not a popular search term, but the results provided on the first page of major search engines are very problematic. This is a classic example of a data void. Outside of white supremacist communities, no one makes websites using this phrase. After the Charleston murders and the discovery of Roof’s manifesto, the news media and other websites covered this issue, diversifying the content available for search engines to associate with that term. At the same time, white nationalists have moved to create and exploit a new data void by shifting their language to “white victims of crimes by blacks,” which continues to driver users to racist and inaccurate content. Terms like these are weaponized and positioned for curious people to stumble upon, whether through Wikipedia (as in Roof’s case), on talk radio, or through countless other venues. Among those seeking to encourage people to rethink their worldview, these “red pills” are seen as invitations to go “down the rabbit hole.”

Like “Harrold-Oklaunion,” few people ever searched for “Sutherland Springs” until November 4, 2017, when news started coming out that a shooter walked into a Baptist church and began shooting. As news rippled out, people turned to search engines to understand what was happening. If we look at searches containing the term “Sutherland Springs,” we can see very few queries before and after the event.



Adversarial actors, intent on capitalizing on what would be a sudden interest in a breaking news story, decided to take advantage of this opportunity. They turned to Twitter and reddit in an effort to associate both the town and, shortly after, the name of the shooter with the term “Antifa.” They knew that there was very little high-quality content about either, which meant that it would not be hard to fill the data void and get the algorithm to rank their content highly in the first hours. Their goal in creating this association was to provide a frame that journalists would have to waste time investigating (to eventually debunk). Furthermore, they wanted early searchers to believe that this shooter’s motives were part of a leftist conspiracy to hurt people. In short order, these adversarial actors managed to influence the front page of search queries, inject “Antifa” into auto-suggest, and trigger journalists to ask whether or not Antifa was involved. In a matter of hours, *Newsweek* ran the headline “‘Antifa’ Responsible for Sutherland Springs Murders, According to Far-Right Media.” This influenced the news content, which search engines take more seriously, and increased the visibility of this association. Furthermore, because of how search engines show headlines, what was shown on the search page was: “‘Antifa’ Responsible for Sutherland Springs Murders...” Eventually, *VICE* ran a headline “No, the Sutherland Springs Shooter Wasn’t Antifa,” and Snopes created a report on the topic. Yet, in a world of data voids, even headlines intended to negate rumors can help spread them.

In addition to data voids that are persistent (e.g., Harrold-Oklaunion and “black on white crimes”) and those that become visible through breaking news, data voids also emerge in auto-suggest. Auto-suggest strives to help users formulate queries more easily by allowing them to quickly select an option as opposed to typing everything out. Many times, this encourages users to expand their search query to add clarity that would benefit the search engine. While search engines restrict auto-suggest from appearing on some terms, the companies do not editorially determine auto-suggest. Rather, the data for auto-suggest comes from previous search queries. The more frequently a particular phrase is used, the more likely it will be suggested in auto-suggest. For example, at the time of writing, Bing’s auto-suggest for “subway” includes the following phrases: menu, specials, coupons, surfers, nutrition, catering, partners, wraps. In this way, it is clear that more people search for this term looking for the sandwich than the mode of transportation. (Clearly, too few people have reliable public transport!)

Journalists and scholars have highlighted deeply disturbing auto-suggestions (e.g., “Jews are...”). These queries are themselves data voids while also producing results that reveal broader data voids. For example, very few people begin search queries with “Jews are...” Those who do often have quite disturbing intentions and are not seeking positive associations with Jewish people. (Recognizing this problem, Bing has added techniques which limit auto-suggest on this phrase.) Auto-suggest is rife with cultural bias, but it is also targeted by those who want to lead people towards data voids. For example, beginning a search query with “black on white...” produces: crime statistics, homicides, crime in America, book, pottery, fabric, violence skyrocketing, and murder stats in auto-suggest.

Most auto-suggest data voids are not the result of adversarial targeting but arise from biases already present in other parts of society. While anti-Semitic hate groups may have tried to manipulate the auto-suggestions associated with “Jews are...”, it is much more likely that these auto-suggestions are produced as a byproduct of prejudicial attitudes among users writing genuine search queries. This does not make this query any less toxic, but it does raise questions about how to best address the caustic query.

Likewise, implicit bias and the amplification of problematic societal attitudes often contribute to the production of data voids even when there’s no maliciously intended actor. For example, image searches for terms like “CEO” tend to return dozens of pictures of white men. This void is likely produced by several overlapping factors. First, the skewed demographics of CEO images are heavily influenced by the similarly skewed demographics of people holding those positions in the “real world.” Especially in Western countries, female CEOs or CEOs of color have been historically underrepresented. Second, when people search for “CEO” in image search, they are in many cases looking for stock photos for a PowerPoint presentation. Getty and other stock photo image producers primarily depict CEOs as white men and only label pictures of white men with the tag of “CEO.” Third, if people who are searching for that term tend to click primarily on pictures of white men, this behavior will reinforce the weighting over time. In this way, a data void rooted in cultural prejudice can be exacerbated from several angles.

## Typology of Data Voids

At this point in our analysis, we have identified three types of voids worth documenting in a growing typology.

- **Data voids that are actively weaponized by adversarial actors immediately following a breaking news event**, usually involving names of locations or suspects in violent attacks (e.g., “Sutherland Springs” or “Parkland.”)
- **Data voids that are actively weaponized by adversarial actors around problematic search terms**, usually with racial, gendered, or other discriminatory intent (e.g., “black on white crime” or “The Greatest Story Never Told” or “white genocide statistics.”)
- **Data voids that passively reflect bias or prejudice in society but are not actively being weaponized** or exploited by a particular group (e.g., “CEO.”)

In addition, we are concerned with search engine behavior in two areas:

- **Auto-suggest** because those suggestions can influence a searcher who might not even have been interested in the suggested language in the first place; and
- **Page 1 search results** because studies show very few people go past the first page.

Contemporary search engines—including Bing and Google—were designed with the goal of allowing users to access information that answer their questions. Unlike social media, which is designed to support the communication and sharing among networks of people, search engines are focused on helping individual seekers find the information they are looking for with few if any limitations. They are consistently evolving to limit the power of those who want to contort the results of search engines for economic, political, or ideological gain.

Data voids are a byproduct of cultural prejudice and a site of significant manipulation by individuals and organizations with nefarious intentions. Addressing data voids cannot be achieved by removing problematic content, not only because removal might go against the goals of search engines but also because doing so would not be effective. Without high-quality content to replace removed content, new malicious content can easily surface. Search engine companies are rightfully wary in limiting what people can search, in no small part because such restrictions on information-seeking raise serious questions about human rights. Finding a way to balance access to information and minimizing the likelihood that people will encounter harmful content is one of the strongest values underpinning contemporary search.

Unlike other forms of content moderation, responding to data voids requires making certain that high-quality content is available in spaces where people may seek to exploit or manipulate users into engaging with malignant information. As search engine companies continue to seek out new ways of improving their algorithmic systems, there is an increasing need for those invested in a healthy internet information ecosystem to think strategically about what kind of content should exist so that fewer people encounter harmful data voids.