Seeking science information online: Data mining Google to better understand the roles of the media and the education system

Elad Segev

Tel Aviv University

Ayelet Baram-Tsabari

Technion, Israel Institute of Technology

Abstract

Which extrinsic cues motivate people to search for science related information? For many science-related search queries, media attention and time during the academic year are highly correlated with changes in information seeking behavior (expressed by changes in the proportion of Google science-related searches). The data mining analysis presented here shows that changes in the volume of searches for general and well-established science terms are strongly linked to the education system. By contrast, ad-hoc events and current concerns were better aligned with media coverage. The interest and ability to independently seek science knowledge in response to current events or concerns is one of the fundamental goals of the science literacy movement. This method provides a mirror of extrapolated behavior and as such can assist researchers in assessing the role of media in shaping science interests, and inform the ways in which lifelong interests in science are manifested in real world situations.
Many interacting factors influence science information seeking on the web. However, people only seek information if they are motivated to fill a gap in their knowledge. This study focuses on the roles of the media and the education system in kindling interest and/or whetting the information needs that lead to seeking science related information on the web. It aims to quantitatively study these roles using data mining of publicly available search query data.

The expanding use of the internet for searching specific science information makes it increasingly possible to use existing web-based tools to learn about users’ information needs and interests (Baram-Tsabari & Segev, 2009). Keywords submitted to search engines can be a useful resource for detecting people’s information needs (Segev & Ahituv, 2010), and may be used to measure issue salience and public agendas without employing survey techniques (Scharrow & Vogelgesang, 2009). For instance, this approach was recently used to study trends in health, economics, and science information seeking (e.g. Anderson, Brossard, & Scheufele, 2010; Baram-Tsabari & Segev, 2009; Choi & Varian, 2009; Ginsberg, Mohebbi, Patel, Brammer, Smolinski, & Brilliant, 2009).

Based on data from Google Trends, Google News, and Google Insights for Search, this study suggests an unobtrusive novel methodology. It uses media attention (expressed by Google News reference volume) and the academic calendar (expressed by the volume of searches for the query 'science') as independent variables explaining changes in information seeking behavior (expressed by changes in the proportion of Google science-related searches).
Seeking science information online

The following literature review will address the concepts of motivation and interest, as well as the ways interest was previously measured in the context of science education; summarize what is already known about science information seeking on the web; and describe ways in which science media impact was traditionally and recently measured.

**Literature Review**

*Motivation, interest and information seeking*

Motivation is commonly understood as the state of wanting to perform a specific activity in a specific situation (Schiefele, 2009). Intrinsic motivation refers to doing something because it is inherently interesting or enjoyable, in contrast to extrinsic motivation, which refers to doing something because it leads to a separable outcome, such as praise or avoidance of punishment (Ryan & Deci, 2000). The Self-Determination Theory proposes that extrinsic motivation can vary greatly in the degree to which it is autonomous (Ryan & Deci, 2000). A student may search online for information about 'particles' only because she fears of her parents/teachers sanctions, or because she personally believes it is valuable for her chosen career. This student may also search for 'particles' out of intrinsic motivation. She was introduced to the concept in class or in a news article and is now interested to know more.

Interest is a powerful motivator (Deci, 1992), which differs from most other motivational concepts by its content specificity (Krapp, 2002). Research on the relationship between interest and learning has traditionally focused on individual and situational interests. Individual interest is considered to be an individual's lasting predisposition to attend to certain stimuli, events, and objects. Situational interest is elicited by certain aspects of the environment. These could be content features, such
as human activity or life themes, or structural features, such as the way in which texts are organized and presented (Ainley, Hidi, & Berndorff, 2002).

Theoretically, situational interest is created in one of two ways: (1) particular factors in the environment focus attention and lead to an affective reaction (e.g. a person watching an engaging documentary about the Large Hadron Collider may be interested in learning more about particles), or (2) an activation of an enduring individual interest (e.g. actualization of a long lasting interest in particles) (Schiefele, 2009). This paper addresses both as one state that may be caused by different factors.

Much research has been devoted to describing students' topics of interests in school science. Students' scientific interests are traditionally identified by questionnaire-based methods which involve asking students to tick boxes in response to a series of prepared questions or topics (e.g. Christidou, 2006; Dawson, 2000; Qualter, 1993; Sjøberg & Schreiner, 2008). Other, more learner-centered research approaches include focus groups (Osborne & Collins, 2001), a student-led review of the science curriculum (Murray & Reiss, 2005), brainstorming sessions, and individual and group interviews (Mcphail, Pierson, Freeman, Goodman, & Ayappa, 2000). Analyzing self-generated science questions is a relatively recent addition to the available research approaches (e.g., Baram-Tsabari et al., 2005, 2006, 2009; Falchetti, Caravita, & Sperduti, 2007). Through studying students' questions, one can learn what students are interested to know about a given topic (Chin & Osborne, 2008). Many of these questions are submitted online to Ask-A-Scientist sites and are mined from the archives by the researchers.
Like questions, search queries\(^1\) reflect a conscious effort to acquire information in response to a need or gap in one's knowledge (Case, 2006). Much like students' questions, science related search queries posted on search engines may be viewed as either indicators of interest (intrinsic motivation) or fulfillment of a requirement (extrinsic motivation).

*Science-related information seeking on the web*

When formal education in science ends, the media become the most available, and sometimes the only source for the public to acquire information about scientific discoveries, controversies, events, and the work of scientists (Nisbet, Scheufele, Shanahan, Moy, Brossard, & Lewenstein, 2002). The print and broadcast media are still the main source of general science and technology information for most Americans (National Science Board, 2008) and Europeans (Eurobarometer, 2007). However, when looking for specific information, most people turn to the internet (Horrigan, 2006).

According to the Pew Internet & American Life Project Poll from December 2009, 74% of American adults (ages 18 and older) use the internet (Rainie, 2010). Roughly 60% of Americans use the internet to get news or information about science and technology, and 20% rely on the internet as their primary source. This is second only to television, which is cited by 41% of Americans as the place where they get most of their science news and information (Horrigan, 2006).

The internet is the source to which people would turn first if they need information on a specific scientific topic. 87% of American users turn to the internet to get some kind of scientific information: look up the meaning of a scientific concept

\(^1\) A term is the basic building block through which a web searcher expresses an information problem when searching on a web search engine. Single or multiple terms and operators form a search query (Spink & Jansen, 2004).
Seeking science information online

or term (70%), learn more about a science story that they heard about offline (65%), complete a science assignment for school (55%), and other science-related information seeking activities (Horrigan, 2006).

Although the internet became the primary source for specific science related information for western adults, research on science-related information seeking is usually conducted in the context of formal schooling. Furthermore, it regularly involves assigned tasks (e.g. Bilal, 2000; Dimopoulos & Asimakopoulos, 2009; MaKinster, Beghetto, & Plucker, 2002), rather than self-generated search tasks (Bilal, 2002). However, what people, including school students, do with online technologies outside the classroom is markedly different from what they do with them in schools, the former usage being more goal-driven, complex, sophisticated, and engaged (Steinkuehler, 2008).

The impact of media science on behavior

There is limited evidence on changes in adult information seeking, retention and use as a result of science coverage on the media. Various studies on media effects suggest that news has a certain impact on science related knowledge (e.g. Miller, Augenbraun, & Kimmel, 2006; Wade & Schramm, 1969). What people know usually corresponds to scientific topics that received the most persistent media coverage (Hargreaves, Lewis, & Speers, 2003). Kahlor and Rosenthal (2009) tested whether antecedents to information seeking as described by their RISP (risk information seeking and processing) model also serve as antecedents to complex and accurate knowledge of global warming. They found that the strongest predictors of knowledge about this topic were education, prior seeking effort, and the number of news media sources used (Kahlor & Rosenthal, 2009).
The introduction and development of data mining techniques have recently made it easier to determine and understand trends in science knowledge seeking. In a recent study Hart and Leiserowitz (2009) collected web traffic data from leading climate change websites, as well as data for media coverage of the topic before and after the release of the blockbuster movie *The Day After Tomorrow*. When *The Day After Tomorrow* was released, global warming related websites had higher levels of web traffic.

The Agenda-Setting model suggests a strong correlation between mass media emphasis on a certain issue and the importance attributed to these issues by the audience (McCombs & Shaw, 1972). This relationship holds under a wide variety of conditions, for a diversity of issues, and when explored by diverse research methods (Dearing & Rogers, 1996).

Halavais (2002) measured changes in word frequency within a large set of popular blogs over a period of four weeks, and compared these changes to those in the traditional media represented on the web. Roberts, Wanta, and Dzwo (2002) operationalized issue salience as the decision of internet users to discuss certain topics in an AOL chat room. Delwiche (2005) used hyperlinks on web logs as behavioral indicators of the perceived importance of an issue as well as an indication of its source. Bentley and Ormerod (2010) further suggested that the ‘adoption’ of ideas could be reflected in people's online searches. Using Twitter as their platform for obtaining crowd-sourced observations, Nagarajan et al. (2009) conceptualized keywords with high search volumes obtained from *Google Insights for Search*\(^2\), as having a greater level of social interest.

---

\(^2\) A description of this tool is provided in the methodology section.
Scharkow and Vogelgesang (2009) argue that Google Insights for Search is a powerful tool for studying agenda setting processes. Compared to survey questions, in search queries there is no interviewer bias or social desirability involved, and the measurement is completely unobtrusive. Moreover, for many users there is virtually no effort involved in using search engines, so that the threshold from issue salience to active information seeking is quite low.

All these examples conceptualize issue salience in terms of observed behavior, rather than internal cognitive or affective change. Similarly, our study operationalized media agenda as the amount of news coverage, while the public agenda was operationalized as the intensity of web searches.

Searches submitted to Google Search have been used to quantitatively characterize science related searches of the general public. For instance, searches on nanotechnology were used to quantitatively characterize science related searches by the general public. Based on the monthly averages of Google Search volume it was found that the public was mostly interested in future directions and applications of nanotechnology, but was less interested in policy and regulatory aspects (Anderson, Brossard, & Scheufele, 2010). Furthermore, when searching with Google for information about nanotechnology, people are likely to encounter health-related content, either through suggested search terms or through the leading search results provided by Google, even for searches not directly related to health. This may create a self-reinforcing spiral that reduces the complexity and variety of information that people are likely to encounter online (Ladwig, Anderson, Brossard, Scheufele, & Shaw, 2010).

Google Search data was also used to identify interests in science and pseudoscience, conduct a cross-national comparison of popular science and
pseudoscience-related searches, and discover possible motivations when searching for specific terms (Baram-Tsabari & Segev, 2009). The findings show that searches for several specific science-related terms, such as 'Global warming', 'Stem cells', 'Intelligent design', and 'Large Hadron Collider' were aligned with the media coverage of those topics. Other searches, such as 'biology' or 'physics' were closely aligned with the academic calendar. The current study further develops and explores the relation between science information seeking behavior, media attention, and the academic calendar.

**Methodology**

*Information sources.* The proportion of internet users searching for answers to specific questions—as opposed to casual browsing—has grown significantly, and as users become more experience online, they increasingly become dependent on search engines (Howard & Massanari, 2007). Google Search provided 65.2% of the online searches in the U.S. in February 2010, followed by Yahoo! Search with 14.1% (Nielsen Online, 2010). Since it is the most widely used, we chose Google and its advanced features as our data source, as it analyzes millions of search queries daily to define the main trends in the *Zeitgeist,* “the spirit of the times” (Google’s definition), with regard to public interest in science. In particular, we used a combination of three tools:

1. *Google Trends (GT)* (www.google.com/trends) first became available to the public in May 2006 to assist research on searches in Google Search and news articles collected in Google News. GT analyzes and displays the proportion of searches for terms, compared to the total number of searches made on Google over a defined period of time (between 2004 and the present). GT also shows how
frequently topics have appeared in *Google News* stories and in which geographic regions people have searched for them the most.

2. *Google Insights for Search (GIS)* ([www.google.com/insights/search/#](http://www.google.com/insights/search/#)) is an offshoot of *GT*. It shows the top searches and searches that experienced significant growth in a specific period and topical category (such as Science, News and Current Events, Entertainment, etc.). It also provides related searches to the original search query and allows for a more advanced cross-national comparison using a visual world map. This tool is still in the beta stage of development.

3. *Google News (GN)* ([news.google.com](http://news.google.com)) is an automated news aggregator available to the public since January 2006 (and in a beta version since 2002). The exact list of news sources is not known outside of *Google*, but *Google* itself reports of over 4,500 English-language news sites, including blogs (Segev, 2008). It aggregates several million articles a day and sends about 1 billion clicks each month to news publishers worldwide (Bharat, 2010), which makes it a reasonable proxy for broader media coverage of news.

We developed a new method to study the correlation between the share of searches for scientific topics (as retrieved from *GT*), their news coverage (as retrieved from *GN*) and the academic calendar (as retrieved by searches for the query ‘Science’ from *GT*) (Figure 1).

[Figure 1 about here]
This involved five steps: (1) Selection of relevant search queries; (2) Data mining; (3) Method evaluation; (4) Method validation; and (5) Using the method to study the source of popular scientific searches.

(1) Selection of relevant search queries

(A) A list of 379 potential English science-related search queries\(^3\) was constructed based on words extracted from various sources, representing a variety of stakeholders: (1) Public interest: popular science search terms on Google were collected from GIS, using the applications of top searches and rising searches\(^4\) (2) Science institutions: The Eurobarometer and Science and Engineering Indicators reports\(^5\) (3) Science and popular science journals: ScienceDaily, New Scientist and Science magazine\(^6\) (4) General media: USA Today, Google News, The Independent, New York Times\(^7\) (5) Aggregating headlines including user generated content: Alltop\(^8\) (6) Science communication research: science terms which are often used in the media that were previously used to assess civic science literacy (Brossard & Shanahan, 2006).

(B) Not all search queries are equally useful. Some are very uncommon, some have more than one meaning, and some are sought after for non-scientific reasons. 'Plasma', for example, is searched frequently for consumer reasons that do not inform us about the interest of the public in science. Because no pre-existing classification scheme could accommodate this unique context, three criteria were developed:

---

3 “Science-related” is used for brevity, but encompasses all areas of science, technology, engineering, and mathematics (STEM). The initial list was constructed on October 15-18, 2009.
4 Rising and top searches for: Large Hadron Collider, cells, science, biology, periodic table, and brain.
5 Terms appearing in the science literacy surveys.
7 USA today: Natural and Physical Sciences selected sub categories; Google news US 2004-2009; The Independent popular topics; NYT most popular alerts.
8 alltop.com
(B₁) Relevance according to category of results. For each search query, GIS provides data regarding the categories to which the results are classified. Science can be the primary category of results (e.g. 'PCR'), the second, and so forth. It may also not be one of the results categories at all (e.g. 'IVF', 'mobile phones', 'space'). Search queries which did not have Science as one of their categories were eliminated, resulting in a list of 301 words.

(B₂) Relevance according to related searches. The level of specificity of a search query was determined using GIS's "related searches" application. This application provides clues to the motivations behind science related searches. Related searches for 'DNA', for example, were mostly content and methodology-related concepts, whereas terms related to the query 'science' were searched for a formal or informal educational-related reason (Baram-Tsabari & Segev, 2009). Search queries that had fewer than 5 out of 10 science related-searches (e.g. 'apes', 'memory') were eliminated from the list, resulting in 259 words. This was the only category that was based on human subjective judgment. Inter-coder reliability was established based on 10% of the sample with a satisfactory agreement.

(B₃) Abundance. The volume of searches for each search query was compared to searches for 'science' using GIS. Search queries with a very low ratio, indicated as "0" by GIS were eliminated from the list (e.g. 'grapheme', 'rationality'), unless over 25% of their results were classified as a 'science' category according to criteria B₁. This procedure resulted in a list of 204

9 Google does not provide a clear explanation of how results are categorized: http://www.google.com/support/insights/bin/answer.py?hl=en&answer=92764 http://www.google.com/support/insights/bin/answer.py?hl=en&answer=141350

10 'science' was selected as a yardstick since it is the name of the category in GIS.

11 The real ratio cannot be 0, since there are enough searches of the word in order for Google to display the data. When very rare search queries are concerned, Google does not provide any data.
words. Examples of classification according to these criteria are presented in Table 1.

[Table 1 about here]

(C) Apart from a mixed group of search queries deriving from various sources (such as public interest, institutions of science and media), we collected three additional homogeneous groups of search queries: (1) 110 search queries were extracted from popular media outlets only\textsuperscript{12}. After screening according to criteria B\textsubscript{1}-B\textsubscript{3} this group resulted in 58 search queries. (2) 113 search queries were extracted from a science education document\textsuperscript{13}. After screening according to criteria B\textsubscript{1}-B\textsubscript{3} this group resulted in 93 search queries. (3) A group of the most popular scientific searches based on data available in GIS was created. GIS provides data regarding the most popular scientific related searches over a period, as well as the top rising searches, i.e. searches that increased their share considerably over a period. Both were collected, resulting in a list of 34 search queries.

(2) Data mining. For each search query data were collected from Google regarding its search volume and news volume. Searches for 'Science' were collected as a measurement for time during the academic calendar. This resulted in three variables:

(A) Searches. Weekly data of searches using Google Search during 2004-2009 were downloaded from GT, resulting in 330 time points expressed on an arbitrary scale.


(B) News coverage. The role of the media was operationalized as the amount of news coverage, and measured as the volume of news results for each search query. A computer program we developed automatically mined from GN monthly data during 2004-2009. This program documented the number of search results returned by GN for each search query limited to a specific month during this time range. For example, GN returned 3,750 results for the search query “DNA” in January 2005. Values varied by month and were expressed as an absolute number. In order to fit the monthly data from GN with the weekly data of GT we used the same value for all the weeks of the same month.

(C) Academic calendar. The role of the education system was operationalized as the time during the academic year, and was measured as searches for the search query 'Science'. A previous study (Baram-Tsabari & Segev, 2009) indicated that searches for 'Science' have a distinctive seasonal trend which is higher during the academic year and lower during the summer and winter vacations. Other search queries, such as 'Biology' and 'Physics' display the same trend. We used searches for ‘Science’ to reflect the role of the education system rather than using the more straightforward school/vacation time, since the search query data are world aggregated. The academic calendar differs between countries and hemispheres and we could not use it as a measure. Additionally, our previous studies show that the share of science-related searches in GT tends to decline over time, since the general variety of searches increases. Searches for ‘Science’ incorporate this trend and make the generation of dummy variables unnecessary. This variable varied by week and was expressed as an arbitrary number. It was identical for all search queries.
(3) **Method evaluation.** A pilot study was conducted in which a pool of scientific related search queries from various sources (news sites, surveys, popular searches, and so on) was divided into two groups ($N_1=92$, $N_2=112$). For each search query a graph was produced using $GT$. One researcher visually examined whether its search trends resembled the news trends, the academic calendar, or both. An expectation for correlation was recorded. Independently, another researcher studied the search queries using our model. Finally, the expected results based on the visual graph were compared with Pearson correlation tests.

(4) **Method validation.** Two groups of search queries were extracted from popular media outlets ($N_3=58$), and a science education document ($N_4=93$). This step was used to test the hypothesis that searches for words collected from media outlets are significantly more correlated with news trends, whereas searches from educational documents are significantly more correlated with the academic calendar.

Additionally, we employed a hierarchical cluster analysis (Aldenderfer & Blashfield, 1984; Johnson, 1967; Lance & Williams, 1967) in order to measure the distances between each pair of search queries in terms of the correlation of their frequency of search with their news coverage and with the academic calendar. Cluster analysis enables grouping together keywords with similar traits. We used Ward's (1963) method, which was found to be the most suitable, as it creates a small number of clusters with relatively more search queries. The Ward method was also proved to outperform the other hierarchical methods (Harrigan, 1985; Punj & Stewart, 1983) in producing homogeneous and interpretable clusters.
(5) *Examination of the role of education and media in motivating the most popular scientific searches.* Popular science-related searches consist of top-searches and rising-searches. While the former tends to be more constant over time and reflects general interest in science, the latter tends to be more ad-hoc and reflects temporal fashions. A group of the most popular scientific searches based on data available in GIS (N₅=34) was used to test the hypothesis that the relatively stable top-searches for scientific issues are related to the academic calendar but that relatively ad-hoc rising-searches are related to the media.

**Results**

In order to assess the roles of the media and the education system in motivating science related information seeking online, the volume of searches was tested for correlations with media attention and the academic calendar. Table 2 presents the percentages of search queries from each group, which were correlated with news coverage, the academic calendar or both.

![Table 2 about here](image)

Science related search queries were more strongly linked to the education system (53%, n = 206) than to news coverage (16.5%, n = 64). An exception was the group of words that was exclusively collected from media outlets (N₃), in which 41% of the search queries were correlated to news trends and 21% to the academic calendar. Only 8.5% of the search queries were significantly correlated both with the academic year and media coverage. Another 22% were not explained by either of these factors.
The results of the method evaluation and validation tests indicate that 70% of the search queries from group 1 and 75% of the search queries from group 2 behaved as expected from $GT$ graphs. Search queries extracted from a science education document were more correlated with the academic calendar whereas those extracted from news outlets were more correlated with news coverage. The differences between news-correlated and academic-correlated searches within the media oriented group were significant ($Z = 1.99$, $p < 0.05$ for $N_3$; $Z = 8.05$, $p < 0.01$ for $N_4$), confirming the first hypothesis (see also Table 2).

Searches with the strongest correlation to news coverage were often related to specific ad-hoc events, such as 'Swine Flu Vaccine', 'West Nile Virus', or specific current concerns, such as 'Greenhouse gas', 'Global Warming' and 'Climate Change'. By contrast, searches with the strongest correlation with the academic calendar were often related to general scientific concepts such as 'Biology', 'Chemistry' and, 'Physics' that play a lasting role in curricula.

Following the hierarchical cluster analysis search queries that had both news and academic orientation could be further divided into four subgroups based on their level of correlation. Figure 2 shows the four subgroups obtained by this analysis. Noticeably, the search queries in each subgroup share a common topic. Subgroup 1 includes environment–related searches such as 'Magnitude', 'Earthquakes' and 'Air Quality', which have a relatively low correlation with both news and the academic calendar ($0.2 < r < 0.5$, $p < 0.01$). Subgroup 2 includes the space-related search for 'Mars Rover', which has a relatively high correlation with news ($r = 0.96$, $p < 0.01$), and a low correlation with the academic calendar ($r = 0.5$, $p < 0.01$). Subgroup 3 includes health–related searches such as 'Cholesterol', 'Eating Disorder' and 'Hearth Disease', which have a relatively low correlation with news ($0.2 < r < 0.5$, $p < 0.01$),
and high correlation with the academic calendar (0.6 < r < 1, p < 0.01). Finally, subgroup 4 includes biotechnology related searches such as 'Biotechnology' and 'Cloning', which have a relatively high correlation with both news (0.4 < r < 0.7, p < 0.01) and the academic calendar (0.8 < r < 1, p < 0.01).

Among the most popular science related search queries (2004-2010), 11 out of 13 were correlated with the academic calendar (Table 3). Most of them were also general scientific concepts as was suggested above. By contrast the top-rising searches were correlated more with news coverage. They included searches related to specific issues and news events such as 'Big Bang'. Other top rising searches, such as 'GCSE Bitesize', were related to new educational portals, and therefore had no correlation with news but rather, to some extent, with the academic calendar. Hence the second hypothesis was partially confirmed.

Discussion

The roles of the media and the education system in motivating science related information seeking online were studied based on the correlation between the volume of searches for a specific search query, its media coverage and time during the academic year.

Since this is a novel analysis approach based on rather new data sources, our first hypothesis was aimed at testing the method using search queries extracted from
well-defined sources. It was assumed that search queries extracted from an educational document would be more aligned with the academic calendar, whereas search queries extracted from news outlets would be better correlated with news coverage. This hypothesis was confirmed. An accidental finding that reinforced the method was the topical clustering of media/academia-oriented searches. Search queries that were related to similar topics (e.g. 'Biotechnology', 'Gene therapy', and 'Cloning'), were plotted in proximity to one another.

Most of the science-related online searches studied here were correlated with the academic calendar rather than with media coverage, pointing to the education system as the origin of the motivation to seek information. This result resonates with findings from the math and science categories on Yahoo! Answers (Adamic, Zhang, Bakshy, & Ackerman, 2008), as well as some of the questions sent to an Ask-A-Scientist website (Baram-Tsabari et al., 2006). Time during the academic year was accountable for almost all the variation in searches for general scientific concepts such as 'Biology', 'Chemistry', and 'Physics', as well as basic concepts such as 'DNA' that are generally, included in secondary science curricula.

The role of the media in motivating science information seeking dominated when search queries were related to a specific ad-hoc event (e.g. 'Mars rover') or current concerns that have not yet been fully incorporated into traditional science curricula (e.g., 'Global warming'). Hart and Leiserowitz (2009) demonstrated that a fictional depiction of a catastrophic event can create a teachable moment of heightened public interest and concern. Here we demonstrated that active pursuit of science knowledge could be the result of media coverage and not only the formal school system. Focused media coverage on current science events and concerns may create a teachable moment, which motivates people to independently search for
related information. This willingness and ability to seek science knowledge in response to current events or concerns is one of the fundamental goals of the science literacy movement.

In order to explore the distinction between media/academia-oriented searches, it was further hypothesized that the top-searches for science related information between 2004 and 2010 would be correlated with the academic calendar, while rising searches (searches that increased their share considerably over a period) during that period would be correlated with media coverage. This hypothesis was confirmed as well.

For some queries the variation in the frequency of searches was almost fully explained by the academic calendar (e.g., 'vector') or the media coverage, but only partially for others (e.g., 'Air quality'). Still other search queries could not be explained by either variable (e.g., 'Moon'). This implies that apart from media coverage and the academic calendar there are other motivations for people to search for scientific information. For example, during the analysis we identified a group of search queries such as 'Full moon', which follow a monthly seasonality. In other words, the motivation of people to look for information on the full moon phenomenon each month does not stem from reading about it on the news or learning about it in school, but rather from their direct experience of this event. We do hope to proceed in this research line, and employ this method to unveil other less obvious motivations for science information seeking and the specific topics attached to those motivations.

It was shown that the main driver for actively seeking science information among British adults is personal circumstances (People Science & Policy Ltd/TNS, 2008). These could not be accounted for using our aggregated data source. However, as long as circumstances are restricted to the individual level and are not widespread
at the social level (e.g., being diagnosed with cancer vs. being diagnosed with an infectious epidemic disease), this variable should not selectively influence searches at the collective level. Moreover, when circumstances do spread to the social level, they are also likely to be discussed in the media.

Bentley and Ormerod (2010) argue that there are two main motivators for searches: (1) rational search by an individual who decides independently to search in response to external events or information, such as news story, and (2) social transmission search when individuals search for a term or topic because other people are searching for it, regardless of topic. In practice individuals will usually be motivated by a combination of the two (Bentley & Ormerod, 2010). Interpersonal influence through social networks (such as Twitter or Facebook) is an important possible motivator for science information seeking that was not specifically included in this study. Although our medium was the internet, our conceptualization of the problem was very traditional: public information seeking is explained by the top-down influence of the media and the formal education system. However, the boundaries between transmitters and receivers of information are much less stable in the ecology of the new media. User generated content was partially accounted for in this research by using GN which aggregates leading blogs. However, a more inclusive measure of interpersonal influence should be included in future studies.

**Study limitations and considerations**

**Causality.** The type of data used in this research constrains the degree to which we could make inferences about the direction of causality between variables based on their observed relationships. In today's blur between new media consumers and producers, it is possible that some media coverage and public searches are
motivated by interests in user generated content as much as by traditional media content.

News sources indexing. Fluctuations in news coverage over time may be an artifact of a particular news source leaving or entering the sample (Stryker, 2008). \( GN \), in particular, does not reveal its full list of sources, which makes it hard for the researcher to control for such events. The list of sources indexed was apparently greatly expanded at the end of 2007. However, at any given time point, the same list of sources was used for all the search queries; therefore, changes in \( GN \) sources affected them evenly.

Lag time. Niederdeppe and Frosch (2009) demonstrated that news coverage influenced short-term consumer purchases of trans-fat products. News effects were strong and significant during the week of their publication and/or a week after they appeared in broadcast or print. The monthly data mined from \( GN \) did not allow for such a distinction between immediate and longer-term effects of media coverage.

Arbitrary scale. Both \( GT \) and \( GIS \) only provide an arbitrary scale system, and not the actual numbers of searches. This makes comparisons regarding the popularity of search queries less straightforward.

Indirect measures. One of this study’s independent variables was an indirect measure of school influence. We did not analyze textbooks or curricula, but used searches for 'Science' as an indirect measure. The second independent variable measured science coverage by the media, but it is impossible to know the extent to which the science news stories were actually read or viewed by those who search with \( Google \).

Behavioral measure. The reliance on aggregate search behavior made it impossible to examine cognitive and affective mediators of news influence. We could
not tell, for example, whether a person searched for 'LHC' because she was interested in high particle physics or annoyed by the way her taxes are spent. By entering a search query people reveal that they are thinking about a topic but we do not know the nature of their thoughts.

The digital divide. Finally, a problem common to all data mining research is the existence of a digital divide. Online research tools represent to some degree the interests of people from industrialized societies, usually from middle and upper class families, who use the Internet’s resources to pursue their science interests.

Notwithstanding these limitations, this study extends the literature by suggesting a new method, which is not subject to the limitations of self-reports, to explore the motivations for scientific information seeking. Some of our main findings—the dominance of education and, to a lesser extent, of media in motivating people to search for scientific information online—are not new to the field of public understanding of science. Yet, they provide a strong support for the validity of our method. Other findings—what search queries are associated with each motivation and their natural tendency for topical clustering—were less obvious. Additionally, the many science-related search queries that were not correlated with media coverage and the academic calendar, such as the searches for seasonal events, open an entirely new dimension for future research using those publicly available tools.

Science-related information seeking is both a means and an end in itself for the science communication and education communities. Individuals who seek science information are more likely to have accurate knowledge about science and to engage with it in their personal and social lives. A greater understanding of the factors that
promote science related information seeking could assist efforts to increase science literacy and public engagement in science.

References


Seeking science information online


Seeking science information online


### Table 1
Examples for classification of search queries

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B1. Relevance</strong></td>
<td>Large Hadron Collider, PCR</td>
<td>DNA, Anatomy, Moon</td>
<td>Pollution, Antibiotics, Cancer genes, Energy</td>
</tr>
<tr>
<td><strong>according to category of results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B2. Relevance</strong></td>
<td>Large Hadron Collider, PCR, DNA, Pollution, Cancer genes, Antibiotics</td>
<td>Moon, Anatomy, Energy</td>
<td>Apes, Memory (Not included)</td>
</tr>
<tr>
<td><strong>according to related searches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B3. Abundance</strong></td>
<td>Cancer genes, Energy, Moon</td>
<td>PCR, DNA, Anatomy, Pollution</td>
<td>Large Hadron Collider, Antibiotics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2  
Percentage of search queries from five groups, which are correlated with news coverage, the academic calendar or both

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>% News &amp; Academic calendar</th>
<th>% News coverage</th>
<th>% Academic calendar</th>
<th>% Not explained</th>
<th>Evaluation and Validation Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pilot A</td>
<td>92</td>
<td>5.4</td>
<td>9.8</td>
<td>56.5</td>
<td>28.3</td>
<td>70%&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Stage 3 - method evaluation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Pilot B</td>
<td>112</td>
<td>19.6</td>
<td>11.6</td>
<td>56.2</td>
<td>12.5</td>
<td>75%&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Stage 3 - method evaluation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Media</td>
<td>58</td>
<td>3.4</td>
<td>41.4</td>
<td>24.1</td>
<td>31</td>
<td>Z=1.99 (p&lt;.05)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Stage 4 - method validation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Educational document</td>
<td>93</td>
<td>4.3</td>
<td>8.6</td>
<td>65.6</td>
<td>21.5</td>
<td>Z=8.05 (p&lt;.01)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Stage 4 - method validation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. GIS Popular searches</td>
<td>34</td>
<td>0</td>
<td>29.4</td>
<td>47.1</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>(Stage 5 - popular searches)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> agreement between expected value according to GT visual graph and the model's results.

<sup>b</sup> difference between number of words correlated with news coverage and number of words correlated with the academic calendar.
Table 3
Pearson correlation of some of the top and rising science-related searches with news coverage and the academic calendar

<table>
<thead>
<tr>
<th>Top Searches</th>
<th>News</th>
<th>Academic</th>
<th>Rising Searches</th>
<th>Query</th>
<th>News</th>
<th>Academic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth</td>
<td>0.36</td>
<td></td>
<td>Big Bang Theory</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td>n.c.</td>
<td>0.95</td>
<td>Big Bang</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>0.94</td>
<td></td>
<td>Dilemme</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>0.94</td>
<td></td>
<td>GCSE Bitesize</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graph</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Periodic Table</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sea</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shark</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Search queries that have a similar meaning (e.g., “GCSE Bitesize”, “BBC Bitesize” and “Bitesize”) were omitted from the table.
Figure 1. Analysis approach

**education system**
- conceptualised: time during the academic year
- measured: searches for 'science' on Google

**media**
- conceptualised: quantity of coverage
- measured: results on Google News

**information seeking**
- conceptualised: information seeking on the web
- measured: searches on Google

**other sources**
- (ineffective, personal experience, interactions between sources, etc)
- not measured
Figure 2. Four groups of search queries scattered based on their Pearson correlation with news coverage and the academic calendar.

*Note:* The Y-axis represents the Pearson correlation between query searches and news coverage. The Y-axis represents the Pearson correlation between query searches and the academic calendar.