Recent work has demonstrated that Web search volume can "predict the present," meaning that it can be used to accurately track outcomes such as unemployment levels, auto and home sales, and disease prevalence in near real time. Here we show that what consumers are searching for online can also predict their collective future behavior days or even weeks in advance. Specifically we use search query volume to forecast the opening weekend box-office revenue for feature films, first-month sales of video games, and the rank of songs on the Billboard Hot 100 chart, finding in all cases that search counts are highly predictive of future outcomes. We also find that search counts generally boost the performance of baseline models fit on other publicly available data, where the boost varies from modest to dramatic, depending on the application in question. Finally, we reexamine previous work on tracking flu trends and show that, perhaps surprisingly, the utility of search in question. The potential predictive power of search activity is illustrated in Fig. 1, which shows the volume of people searching for the movie Transformers 2 (Fig. 1A) and the video game Tom Clancy's H.A.W.X. (Fig. 1B) around their release dates. In both cases searches peak either on or slightly after the release date, but precursors to the spikes appear several days or even weeks in advance. Correspondingly for music, Fig. 1C shows that the rank


The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

†To whom correspondence may be addressed. E-mail: goel@yahoo-inc.com, hofman@yahoo-inc.com, lahaies@yahoo-inc.com, pennockd@yahoo-inc.com, or djw@yahoo-inc.com.
of the song “Right Round” in terms of search volume closely tracks its rank on the Billboard Hot 100 chart.

Thus motivated, we now investigate whether search activity is a systematic leading indicator of consumer activity by forecasting (i) opening weekend box-office revenue for 119 feature films released in the United States between October 2008 and September 2009; (ii) first-month sales of video games across all gaming platforms (e.g., Xbox, PlayStation, etc.) for 106 games released between September 2008 and September 2009; and (iii) the weekly rank of 307 songs that appeared on the Billboard Hot 100 list between March and September 2009. Search data for movies and video games come from Yahoo!’s Web search query logs for the US market. Predictions in these domains are based on linear models with Gaussian error of the form

$$\log(\text{revenue}) = \beta_0 + \beta_1 \log(\text{search}) + \epsilon,$$

where, in order to account for the highly skewed distributions of popularity, both revenue and search volume are log-transformed. For songs, search data were collected from Yahoo!’s dedicated music site, music.yahoo.com. We predict the weekly Billboard rank using search rank from the current and previous weeks:

$$\text{billboard}_{t+1} = \beta_0 + \beta_1 \text{search}_t + \beta_2 \text{search}_{t-1} + \epsilon.$$

Fig. 2 A–C shows that search-based predictions are strongly correlated with realized outcomes for movies (0.85) and video games (0.76) and moderately correlated for music (0.56), where in each case revenue or rank is predicted on the day immediately preceding the event of interest. Moreover, Fig. 2 D–F shows that the predictive power of search persists as far out as several weeks in advance—for example, four weeks prior to a movie’s release.
search volume remains highly correlated (0.75) with opening weekend revenue. Going beyond correlation with contemporaneous events, therefore, these results show that search can also predict the near future—a finding that may apply usefully to a wide range of consumer behaviors (e.g., airline travel, hotel vacancy rates, and auto sales) and economic indicators (e.g., real-estate prices, credit card defaults, and consumer confidence indices).

To put these results in the proper perspective, however, we next compare all search-based predictions with simple models built on publicly available information. For movies, we generate baseline predictions using a linear model that includes production budgets, the number of screens on which each movie opened, and box-office projections from the Hollywood Stock Exchange (HSX) (hsx.com), an online, play-money prediction market that is known to generate informative predictions (12):

$$
\log(\text{revenue}) = \beta_0 + \beta_1 \log(\text{budget}) + \beta_2 \log(\text{screens}) + \beta_3 \log(\text{HSX}) + \epsilon.
$$

For video games, many of the key indicators of revenue, including production budgets and initial game supply, are not publicly available. Therefore we create baseline predictions using critic ratings on a scale from 1 to 10:

$$
\log(\text{revenue}) = \beta_0 + \beta_1 \text{rating} + \epsilon.
$$

For games that are sequels—a category that includes many of the most anticipated and profitable video games (e.g., Final Fantasy XIII)—we also include in the model the lifetime revenue of the game’s immediate predecessor:

$$
\log(\text{revenue}) = \beta_0 + \beta_1 \text{rating} + \beta_2 \log(\text{predecessor}) + \epsilon.
$$

where we evaluate predictions for sequel and nonsequel games separately. Finally, for music we generate baseline predictions from an autoregressive model:

$$
\text{billboard}_{t+1} = \beta_0 + \beta_1 \text{billboard}_{t-1} + \epsilon,
$$

where the reporting lag associated with published Billboard rankings requires that we predict outcomes for week $t+1$ using data from week $t-1$.

Fig. 3 A–C shows that the baseline models outperform the search-based predictions for movies (correlation 0.94), music (0.70), and sequel video games (0.80). However, when we consider nonsequel video games—in which case baseline predictions are based only on critic ratings—the correlation between baseline and actual outcomes drops to 0.44, compared to 0.77 for search-based predictions. These results illustrate two points: first, although search data are indeed predictive of future outcomes, alternative information sources often perform equally well or even better; second, search appears to be most useful when key indicators (e.g., past sales performance, production budgets, etc.) do not exist or are unavailable.

Next, we consider the performance of combined models that incorporate both search and baseline data. For example, to predict movie revenue we fit a linear model that includes search data along with production budgets, opening screens, and HSX estimates. We find that the magnitude of improvement over the baseline models varies greatly (Fig. 3 D–F). The performance of the augmented model for movies is virtually identical to that of the baseline model, and for video game sequels, the performance improves only modestly, from 0.80 to 0.83. For music, however, performance increases from 0.70 for the baseline to 0.87 for the augmented model—a much larger improvement—and for nonsequel video games, the corresponding improvement is dramatic, increasing from 0.44 to 0.80.

![Fig. 3](image-url)
Given the attention that search-based predictions have received recently, it may seem surprising that search data are, at least in some cases, no more informative than traditional data sources. To further examine this point, we reassess the utility of search data in monitoring influenza caseloads, arguably the best-known example of search queries correlating with real-world outcomes. To place that work in context, we note that public reports of flu caseloads by the Centers for Disease Control and Prevention (CDC) are typically delayed one to two weeks; thus the primary aim of search-based flu estimation is real-time monitoring, as opposed to predicting future activity. In particular, flu trend estimates for a given week rely on query volume during that same week. At the end of week $t$, query volume from that week is used to estimate the yet-to-be-reported flu level for week $t$. To better gauge the value of search data for flu surveillance, we compare search-based estimates to those from a simple autoregressive model:

$$\text{flu}_t = \beta_0 + \beta_1 \text{flu}_{t-2} + \beta_2 \text{flu}_{t-3} + \epsilon.$$  

We find a correlation of 0.86 between current flu levels and recently reported levels where, because of a two-week reporting delay, the estimated flu level during week $t$ is based only on reported levels from weeks $t-2$ and $t-3$ (Fig. 4). By comparison, Ginsberg et al. (10) showed that search activity in a given week for terms such as “flu” and “cold” has a correlation of 0.94 with the number of influenza caseloads that week. Search, therefore, does indeed outperform the autoregressive model; however, the difference in performance is relatively small. Moreover, we note that the CDC has recently deployed a system to reduce reporting delays. To illustrate how such an improvement would impact the relative performance of search-based prediction, we repeat the comparison under the assumption that flu reports are delayed by only one week instead of two. In this case, the corresponding autoregressive model has a fit of 0.95, nearly identical to that of the search-based estimates (0.94) of Ginsberg et al. (10). We conclude, therefore, that in monitoring the flu—as in some of the other domains we have considered—search data are comparable in utility to alternative information sources, but not necessarily superior.

**Discussion**

Across the empirical cases that we have considered, we observe that search counts are generally predictive of consumer activities, such as attending movies and purchasing music or video games, that will take place days or even weeks in the future. As Fig. 5 shows, however, there is also wide variability in the predictive power of search among the different domains. The reasons for these intriguing differences are ultimately unclear and suggest the need for further work; however, we offer the following three possible factors.

First, the size of the relevant population varies greatly across domains. A major feature film, for example, may draw tens of millions of viewers within the United States, whereas popular, platinum albums sell on the order of 1 million copies. One might expect, therefore, some variation in signal strength to be accounted for by variation in overall market size across domains. Second, in certain domains searching may be more closely tied to the measured outcomes than in others. For example, it may be the case that movie searches are primarily intended to locate theaters or to purchase tickets online—activities directly related to box-office revenue—whereas music searches may simply reflect interest in a song’s lyrics, as opposed to any intent to purchase it. Finally, the difficulty in identifying relevant queries varies considerably among domains. With movies or video games, for example, one would expect users to search by the name of the object in question, whereas with music it is frequently the case that the song name is less well known than the name of the artist. Because artist searches may arise for many reasons, and may also be conflated with names of other individuals, it is arguably harder to classify these queries. It was for this reason, in fact, that we chose to study song-related search on Yahoo! Music rather than by using general Web search data, in part mitigating the problem of identifying relevant queries.

In addition to variation in the predictive power of search alone, Fig. 5 also indicates substantial differences in the relative value of search data compared to alternative sources. For movies, search is clearly outperformed by the baseline and offers no improvement to it in the combined model. For video games, meanwhile, search performs well relative to the baseline for nonsequels and poorly for sequels, but in each case boosts the baseline in the combined model. And for music, search performs worse than the baseline, but offers a substantial boost in the full model. These findings raise interesting questions regarding the circumstances under which search-based predictions might be useful. Most obviously, the importance of the baseline models suggests that search-based predictions can yield the greatest performance boost when key information—such as marketing or production budgets—is difficult to acquire. In the case of nonsequel video games, for example, it is likely that insiders (e.g., employees at video game companies) would have access to better baseline information and would therefore gain less from search counts than outsiders who lack such information.

Even in domains where baseline models generally perform well, however, search data may still help in some circumstances. In the case of music, for example, although the autoregressive model in general outperforms search, the model performs particularly poorly for highly ranked songs, which have a tendency to drop off rapidly, as can be seen in Fig. 3C. Because search counts

---

**Fig. 5.** The correlation between predicted and actual outcomes for movies, video game sequels and nonsequels, music, and flu.
help compensate for this weakness of the autoregressive model, they boost its performance in the combined model. Likewise, sudden changes in search volume may help identify so-called “turning points” in economic time series, where otherwise accurate baseline models (which are usually autoregressive in nature) tend to fail badly (4, 13).

Finally, we note two further points that suggest the potential value of search-based predictions. First, modest performance gains may still prove useful for applications, such as financial analysis, where even a minimal performance edge can be valuable. Second, unlike other data sources that require customized and often cumbersome collection strategies, search data can be collected for many domains simultaneously and easily analyzed along geographic and other dimensions, all in real-time. Ultimately, therefore, the utility of search counts for predicting real-world events may have less to do with their superiority over other data sources than with matters of speed, convenience, and flexibility across a variety of domains.

Materials and Methods

To identify user intent from queries we applied a simple and effective heuristic that leverages search engine technology. User queries were categorized as movie-related if an Internet Movie Database (IMDb) link appeared in the first page of search results. We mapped queries to specific movies by extracting the unique movie identifier in the corresponding IMDb link; when multiple IMDb links appeared, we determined user intent from the top-ranked result. For example, “transformers”, “revenge of the fallen”, and common misspellings of these queries all return the result URL www.imdb.com/title/tt0553697, a link to the IMDb page for Transformers 2: Revenge of the Fallen. Although it is likely that our method will conflate searches for movies with common-word names such as “transformers” with other unrelated searches (e.g., for transformers of the electrical equipment variety), we expect that these other searches will be uncorrelated with the release of the movies in question; thus variations in search volume should still be informative. Analogous to our analysis of movies, we mapped search queries to video games by extracting game-specific identifiers from URLs for three leading gaming Web sites: GameTrailers, GameSpot, and GamePro. In contrast to our analysis of movies and video games, search volume for music is calculated by tallying queries on Yahoo! Music that contain a song title, although it is likely that our method will conflate searches for songs with common-word names such as “transformers” with other related searches (e.g., for transformers of the electrical equipment variety), we expect that these other searches will be uncorrelated with the release of the songs in question; thus variations in search volume should still be informative.

We thank Jitendra Waral, Vanessa Colella, and Nick Weir for helpful conversations.

ACKNOWLEDGMENTS. We thank Jitendra Waral, Vanessa Colella, and Nick Weir for helpful conversations.