Predicting consumer behavior with Web search

Sharad Goel[1], Jake M. Hofman[1], Sébastien Lahaie[1], David M. Pennock[1], and Duncan J. Watts[1]

Microeconomics and Social Systems, Yahoo! Research, 111 West 40th Street, New York, NY 10018

Edited* by Simon A. Levin, Princeton University, Princeton, NJ, and approved August 10, 2010 (received for review April 29, 2010)

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 data relative to a simple autoregressive model is modest. We conclude that in the absence of other data sources, or where small improvements in predictive performance are material, search queries provide a useful guide to the near future.

culture | predictions

As people increasingly turn to the Internet for news, information, and research purposes, it is tempting to view online activity at any moment in time as a snapshot of the collective consciousness, reflecting the instantaneous interests, concerns, and intentions of the global population (1, 2). From this perspective, it is a short step to conclude that what people are searching for today is predictive of what they will do in the near future. Consumers contemplating buying a new camera may search to compare models; moviegoers may search to determine the opening date of a new film, or to locate cinemas showing it; and individuals planning a vacation may search for places of interest, to find airline tickets, or to price hotel rooms. If so, it follows that by appropriately aggregating counts of search queries related to retail activity, moviegoing, or travel, one might be able to predict collective behavior of economic, cultural, or political interest. Determining the nature of behavior that can be predicted using search, the accuracy of such predictions, and the time scale over which predictions can be usefully made are therefore all questions of interest.

Although previous work has considered the relation between search volume and offline outcomes, researchers have focused on the observation that search "predicts the present" (3, 4), meaning that search volume correlates with contemporaneous events. For example, Ettredge et al. (5) found that counts of the top 300 search terms during 2001–2003 were correlated with US Bureau of Labor Statistics unemployment figures; Cooper et al. (6) found that search activity for specific cancers during 2001–2003 correlated with their estimated incidence; and Eysenbach (7) found a high correlation between clicks on sponsored search results for flu-related keywords and epidemiological data from the 2004–2005 Canadian flu season. More recently, Polgreen et al. (8) showed that search volume for handpicked influenza-related queries was correlated with subsequently reported caseloads over the period 2004–2008, and Hulth et al. (9) found similar results in a study of search queries submitted on a Swedish medical Web site. An automated procedure for identifying informative queries is described in Ginsberg et al. (10), and based on that methodology, Google Flu Trends (http://www.google.org/flutrends) provides real-time estimates of flu incidence in several countries. Finally, Choi and Varian (3, 4) have compared search volume to economic activity, including auto and home sales, international visitor statistics, and US unemployment claims; and similar work has been reported for German unemployment claims (11). In this paper, we further this work by considering the ability of search to predict events days or weeks in advance of their actual occurrence.

In so doing we also emphasize an often overlooked aspect of prediction—namely, that performance is relative. To illustrate, consider predicting the weather in Santa Fe, New Mexico, where it is sunny 300 days a year. A prediction of sunshine every day would be correct 82% of the time, yet hardly impressive; nor could a model that fails to outperform the simple, autoregressive rule that tomorrow's outcome will be like today's be said to be predictive in any interesting way. Correspondingly, the predictive power of search should be judged in relation to statistical models fit with traditional data sources, prediction markets, or expert opinions—a point that, with some exceptions (3, 4), has been overlooked in previous related work.

Finally, we extend the domain of past studies from epidemiological and macroeconomic time series to include consumer activity such as that associated with movies, music, and video games. These are a natural class of events to consider as they represent activities (e.g., attending a movie) for which it is plausible that individuals might (*i*) harbor the intention to perform the corresponding action sometime in advance of actually fulfilling it and (*ii*) signal that intention through a related Web search. In this sense, our paper is related to other work that uses online chatter and Twitter posts to predict rankings of books on Amazon.com (1) and movie box-office revenues (2), respectively. However, by studying three classes of outcomes—the opening weekend box-office ticket sales for feature films, the first-month revenues of video games, and the weekly ranks of songs on the Billboard Hot 100 chart—we reveal some additional insights regarding variation across domains in how search-based predictions perform, both in an absolute sense and relative to alternative forecasting methods.

Results

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. 1*A*) and the video game *Tom Clancy's H.A.W.X.* (Fig. 1*B*) 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. 1*C* shows that the rank of the song "Right Round" in terms of search volume closely tracks its rank on the Billboard Hot 100 chart.

Author contributions: S.G., J.M.H., S.L., D.M.P., and D.J.W. designed research; S.G., J.M.H., and S.L. performed research; S.G., J.M.H., and S.L. analyzed data; and S.G., J.M.H., S.L., D.M.P., and D.J.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]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.

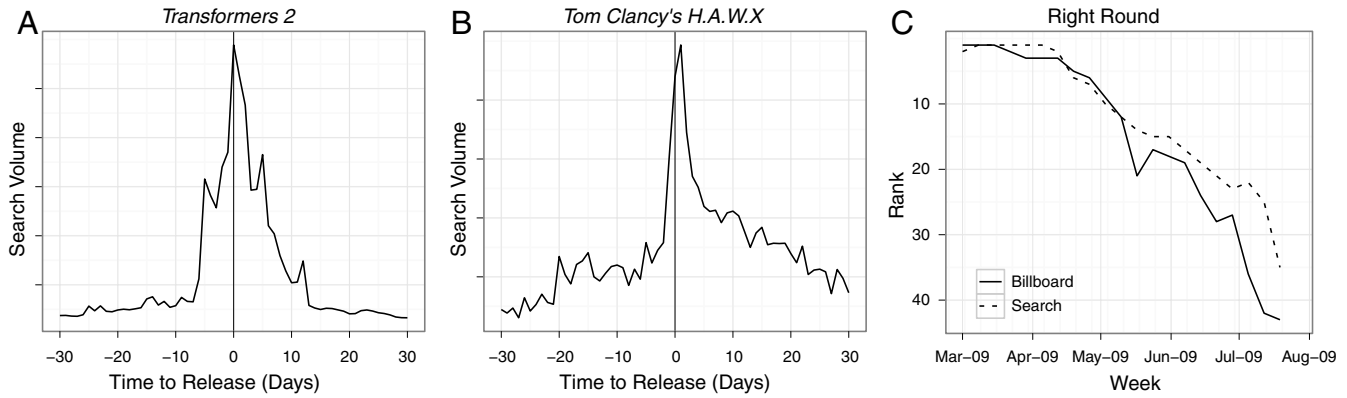

Fig. 1. Search volume for the movie *Transformers 2* (*A*) and the video game *Tom Clancy's H.A.W.X.* (*B*) prior to and after their release, and search and Billboard rank for the song "Right Round" by Flo Rida (*C*).

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

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

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:

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

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

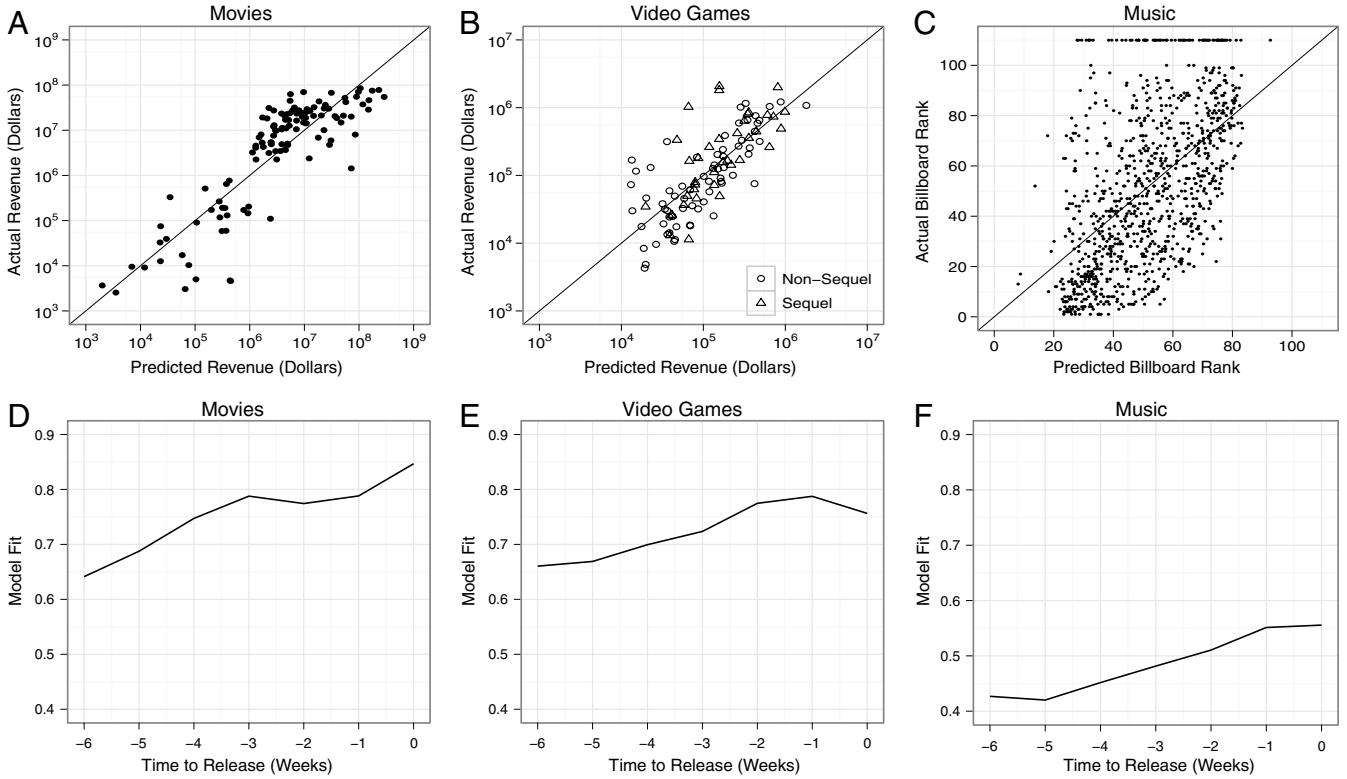

Fig. 2. Search-based predictions for box-office movie revenue (*A*), first-month video game sales (*B*), and the Billboard rank of songs (*C*), where predictions are made immediately prior to the event of interest; correlation between predicted and actual outcomes when predictions are based on query data *t* weeks prior to the event (*D–F*).

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):

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

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:

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

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:

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

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

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

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.

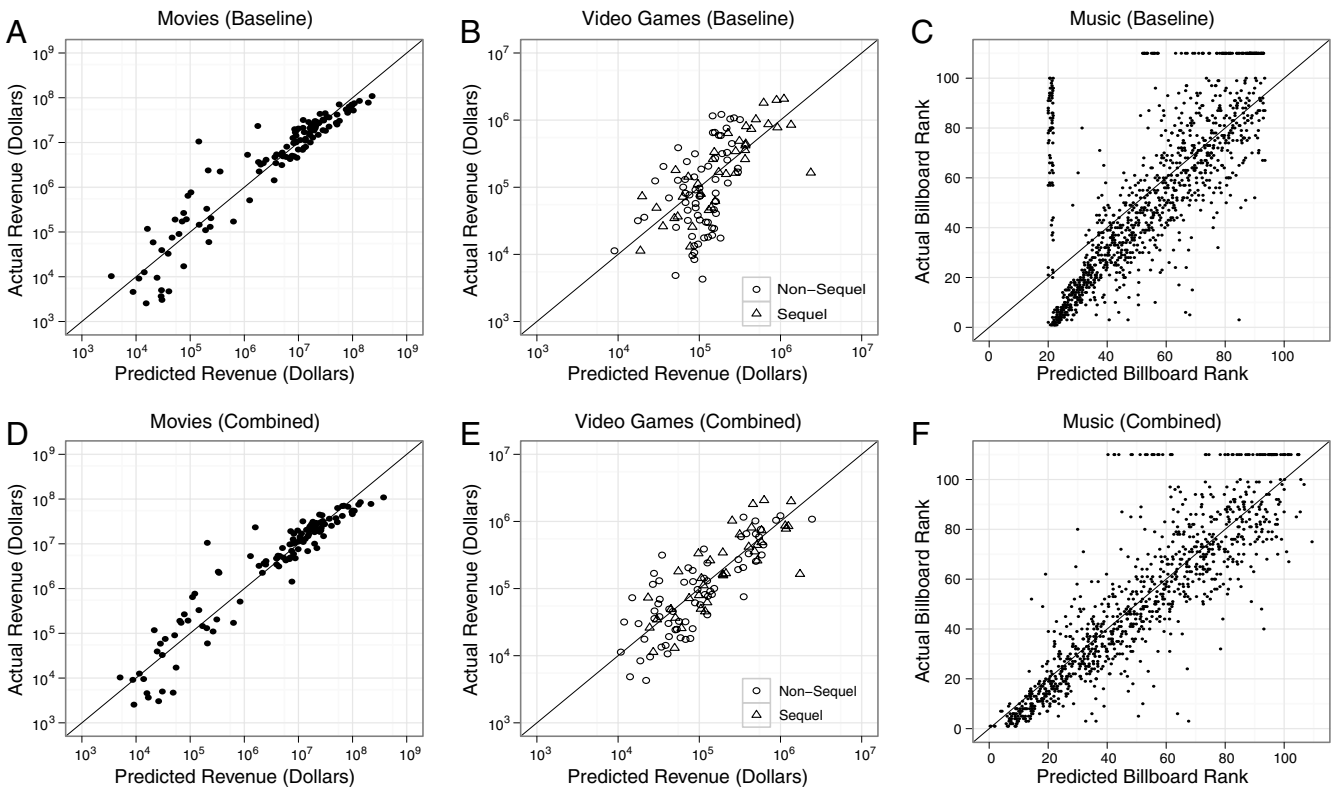

Fig. 3. Predictions from the baseline (*A–C*) and the combined baseline-plus-search models (*D–F*) for movies, video games, and music.

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:

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

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

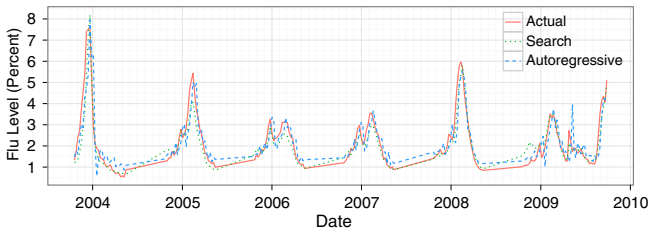

Fig. 4. Actual and estimated flu levels in the United States, where flu level is the percentage of physician visits that involve patients with influenza-like illnesses. Search-based estimates are from Google Flu Trends.

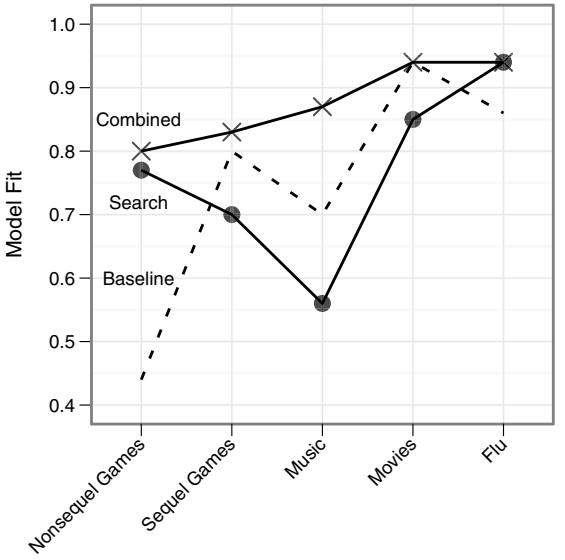

Fig. 5. The correlation between predicted and actual outcomes for movies, video game sequels and nonsequels, music, and flu.

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. 3*C*. Because search counts

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/tt1055369/, 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's (normalized) title. Restricting to vertical search on Yahoo! Music enables us to extract relevant signal for songs with succinct and common titles (e.g., "Then" by Brad Paisley) for which intent in general Web search is harder to discern.

Movie data on revenue, budget, and number of opening screens were obtained from IMDb. Revenue ranged from \$3K to \$109M, with a mean of \$17M and a median of \$10M, and budgets ranged from \$1.5M to \$250M, with a mean of \$44M and a median of \$24M. The number of opening screens was moderately correlated with budget (0.6) and ranged from 1 to 4,325, with a mean of 2,074 and a median of 2,507.

Although it would have been interesting to include the marketing budget for movies in the baseline model, this information was not publicly available. However, we suspect that much of the information contained in the marketing budget would also be reflected in the number of screens chosen by the distributors, as both variables likely reflect insider expectations of success, and both may also create self-fulfilling prophecies (i.e., because widely distributed and promoted films may attract viewers solely on the strength of their distribution and promotion).

Sales and critic ratings data for video games were obtained from VGChartz (vgchartz.com), a leading video game sales tracking Web site. First-month sales ranged from \$4K to \$2M, with a mean of \$2.9M and a median of \$1.1M. Ratings ranged from 3.1 to 9.5, with a mean of 7.5 and a median of 7.7. Lifetime revenue for the 38 sequels in our dataset were also obtained from VGChartz.

Finally, the music data are based on the Hot 100 chart, which is issued weekly by the magazine *Billboard*, and lists the 100 most popular songs in the United States based on airtime and sales. Chart rankings were acquired through the Billboard Developer API (developer.billboard.com) and include artist, song title, rank, and chart release date. There is a substantial reporting delay in the Billboard charts, with the currently available chart providing information collected over a week ago. In our analysis we assume access only to information publicly available at the time of prediction; thus we explicitly abstain from predicting when a song first enters the Billboard charts. Instead, for any given week we predict the future rank of songs appearing on the currently available Billboard Hot 100 chart.

To guard against overfitting our models to the data, movie and video game predictions were generated via leave-one-out estimation. For example, a prediction for each of the $n = 119$ movies was first generated by a model trained on the other $n - 1$ movies, after which model performance was evaluated. In the case of music, for each of the 13 weeks between June and August 2009, we trained our models on the previous three months of data and evaluated the following week's predictions.

Flu incidence levels were obtained from the CDC, and search-based estimates were taken from www.google.org/flutrends.

Data aggregation was carried out with Hadoop and Pig (14), statistical analysis was conducted in R, and graphs were generated with ggplot2 (15).

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

1. Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM, New York), pp 78–87.
2. Asur S, Huberman B (2010) *Predicting the Future with Social Media* arXiv:1003.5699.
3. Choi H, Varian H (2009) Predicting initial claims for unemployment benefits. Available at http://research.google.com/archive/papers/initialclaimsUS.pdf.
4. Choi H, Varian H (2009) Predicting the present with Google Trends. Available at http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.
5. Ettredge M, Gerdes J, Karuga G (2005) Using web-based search data to predict macroeconomic statistics. *Commun ACM* 48:87–92.
6. Cooper C, Mallon K, Leadbetter S, Pollack L, Peipins L (2005) Cancer Internet search activity on a major search engine, United States 2001–2003. *J Med Internet Res* 7(3):e36.
7. Eysenbach G (2006) Infodemiology: Tracking flu-related searches on the Web for syndromic surveillance. *American Medical Informatics Association Annual Symposium Proceedings* (Curran Associates, Red Hook, NY), pp 244–248.
8. Polgreen PM, Chen Y, Pennock DM, Nelson FD (2008) Using Internet searches for influenza surveillance. *Clin Infect Dis* 47:1443–1448.
9. Hulth A, Rydevik G, Linde A (2009) Web queries as a source for syndromic surveillance. *PLoS ONE* 4(2):e4378.
10. Ginsberg J, et al. (2008) Detecting influenza epidemics using search engine query data. *Nature* 457:1012–1014.
11. Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. *Appl Econ Quart* 55:107–120.
12. Wolfers J, Zitzewitz E (2004) Prediction markets. *J Econ Perspect* 18:107–126.
13. Kling JL (1987) Predicting the turning points of business and economic time series. *J Bus* 60:201–238.
14. Olston C C, Reed B, Srivastava U, Kumar R, Tomkins A (2008) Pig latin: A not-so-foreign language for data processing. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (ACM, New York), pp 1099–1110.
15. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York).