What Can Search Predict?

Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, Duncan J. Watts
Yahoo! Research, 111 West 40th Street, New York, NY 10018, USA
{goel, hofman, lahaies, pennockd, djw}@yahoo-inc.com

ABSTRACT

Recent work has shown that search query volume correlates well with a variety of phenomena, from influenza caseloads to economic indicators like real-estate prices, auto sales, and travel statistics. In this paper, we investigate the degree to which search behavior predicts the commercial success of cultural products, namely movies, video games, and songs. In contrast with previous work that has focused on real-time reporting of current trends, we emphasize that here our objective is to predict future activity, typically days to weeks in advance. Specifically, we use query volume to forecast 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. In all cases that we consider, we find that search counts are indicative of future outcomes, but when compared with baseline models trained on publicly available data, the performance boost associated with search counts is generally modest—a pattern that, as we show, also applies to previous work on tracking flu trends. We conclude that in the absence of other data sources, or where small improvements in predictive performance are material, search queries may provide a useful guide to the near future.

Keywords

cultural events, forecasting, web search

1. INTRODUCTION

As people increasingly turn to web search engines for news, information, and research purposes, it is tempting to view the corpus of search queries at any moment in time as a snapshot of the collective consciousness, reflecting the instantaneous interests, concerns, and intentions of the global population of Internet users. 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; movie goers 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, movie going, 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.

In this paper, we address these issues in the context of cultural events such as those 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 (a) harbor the intention to perform the corresponding action some time in advance of actually fulfilling it; and (b) signal that intention through a related web search. By this standard, almost any kind of consumer-based economic activity would seem reasonable; however, we focus here on three prediction tasks: 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.

In studying these three specific empirical cases, we make two contributions to the emerging body of work relating web search to real-world outcomes.

First, whereas previous work has emphasized the use of search counts to perform real-time reporting, or what has been called "predicting the present" [2, 3], here we are specifically interested in predicting future events. Past work, for example, has shown that influenza caseloads in a given week correlate well with query volume over the same period [8, 13]. As the conventional monitoring of influenza typically involves a reporting delay of 1–2 weeks, the ability to tally query volume, and hence estimate incidence, in near real-time has the potential to aid public health surveillance. These more timely estimates, however, are not strictly speaking predictions of future activity. By contrast, we find that for the domains we consider, search by itself is a reasonable guide for predicting outcomes ranging from a few days to a few weeks in the future.

Second, in assessing how well search counts predict real-world outcomes, we emphasize an obvious but nevertheless often overlooked aspect of prediction performance, namely that all performance is relative. To illustrate, consider the exercise of predicting tomorrow's weather in Santa Fe, NM. Santa Fe records on average 300 days of sunshine per year, so a minimally informative model that predicts sunshine every day would still be correct 82% of the time. More generally,

Paper available at `http://cam.cornell.edu/~sharad`.

a model that fails to outperform the simple, autoregressive rule that tomorrow's weather will be like today's cannot be said to be predictive in any interesting way, no matter how highly correlated its predictions are with reality. What these examples highlight is that performance is meaningful only in relation to some baseline means of prediction, such as statistical models, prediction markets, or expert opinions; yet with the exception of the working papers of Choi and Varian [2, 3], past work does not generally compare the performance of search-based predictions to baselines. By contrast, we explicitly quantify the performance of search relative to models trained on publicly available data. As we show, both for our own empirical cases and also for the previously studied case of influenza, search data typically provide only a modest boost to the performance of such baselines.

The remainder of this paper proceeds as follows. In the next section we briefly outline related work. In Sections 3.1–3.3, we describe results for the three primary domains that we consider: movies, video games, and music, respectively. For each of these cases, we describe (a) the method for inferring user intent from search queries; (b) the publicly available data from which we construct corresponding baseline models; (c) the specification of the baseline, search, and combined prediction models; and (d) the relative performance of these models. In Section 4, we further discuss the issue of relative performance by revisiting the example of predicting influenza caseloads from query volume, showing that previously published search-based estimates are comparable to those from a simple autoregressive model. And in Section 5 we conclude, discussing some remaining unresolved issues as well as directions for future research.

2. RELATED WORK

The use of search volume as an indicator of real-world behavior and events is a problem that has only been addressed relatively recently in the research literature. In early work, Ettredge et al. [6] found that counts of the top 300 search terms during 2001–2003 were correlated with U.S. Bureau of Labor Statistics unemployment figures; Cooper et al. [4] 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. [13] showed that normalized counts of hand-picked influenza-related search queries were 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. [8], and based on that methodology, Google Flu Trends (`google.org/flutrends`) provides real-time estimates of flu incidence in several countries.

In addition to work that focuses specifically on web search, a growing body of research examines web-related behavior in general as an indicator of other kinds of activity. For example, Corley et al. [5] have shown that influenza-related blog posts were correlated with the onset of the U.S. 2008 flu season; and Wilson and Brownstein [14] have combined search queries with volume of news articles to detect an outbreak of the food-borne bacteria listeriosis in Canada.

Most closely related to the current work are two working papers by Choi and Varian [2, 3]. The first of these finds that the addition of normalized search counts to baseline, autoregressive models of economic time series (e.g., auto and home sales, and international visitor statistics) improves model fit, generally by small amounts [3]. Subsequently, the same authors have described similar results for initial claims for U.S. unemployment benefits [2]—a result that has also been reported for Germany [1].

3. PREDICTING THE SUCCESS OF CULTURAL PRODUCTS

Below we describe our findings for movies, video games and music.

3.1 Movies

Data Description. We forecasted opening weekend box-office revenue for 119 feature films released in the U.S. between October 2008 and September 2009. Revenue data were obtained from the Internet Movie Database (IMDB), and ranged from $3K to $109M, with a mean of $17M and a median of $10M. To construct a baseline prediction model, we incorporated film production budgets, the number of screens on which each movie opened, and box office projections from Hollywood Stock Exchange (`hsx.com`), an online, play-money prediction market that specializes in entertainment events, and that is known to generate informative predictions [15]. As with the revenue data, budget and screen information was obtained from IMDB. 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 4325, with a mean of 2074 and a median of 2507.

Daily search volume for movies was based on Yahoo!'s web search query logs for the U.S. market, and spans the period September 2008 to September 2009. Identifying user intent from queries was a recurring challenge in this work. In the case of movies, we applied a simple and effective heuristic that leverages the search engine algorithm. User queries were categorized as movie-related if an 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*.

This approach to query categorization generally captures user intent. Figure 1 plots query volume for *Transformers 2* during the one month period before and after the movie's release, where one observes a sharp spike in volume around its June 24, 2009 release date. We note, however, that some queries may lead this method astray. For example, though the query "fandango" returns as a top result an IMDB link to the 1985 movie of the same name, most users are likely searching for Fandango.com, the popular website for purchasing movie tickets. Such instances of ambiguous user intent appear to be rare, and, at least for the movies we consider, do not appreciably affect our results.

Analysis. To predict opening weekend revenue, we fit simple linear models after log-transforming the inputs and outputs. We consider two standard measures of model fitness: the correlation, and the mean absolute error (MAE)

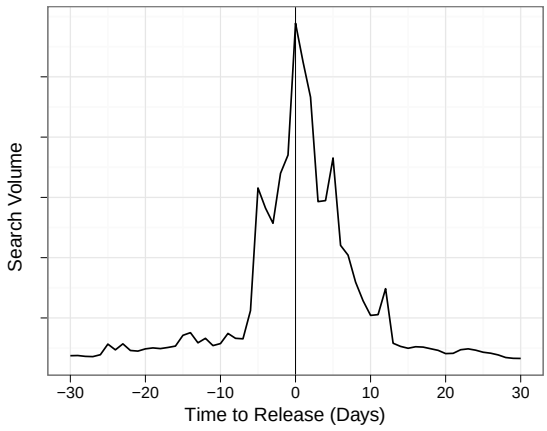


Figure 1: Query volume for the movie *Transformers 2* leading up to and after its release.

between predicted and actual (log-transformed) outcomes. In both cases, fitness is computed via leave-one-out cross-validation. That is, 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.

We start by investigating the performance of search alone in predicting movie revenue. Specifically, forecasts are generated via the model:

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

As shown in Figure 2, search exhibits good prediction performance several weeks in advance of a movie's opening. In particular, when predictions are based on total query volume during a one-week period one month prior to a movie's release, we find a correlation of 0.75 between predicted and actual outcomes. Moreover, when predictions are based on query volume during the 7 days immediately prior to release, this correlation increases to 0.85.

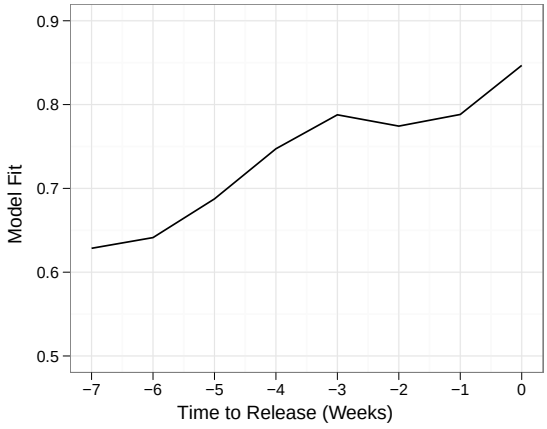


Figure 2: Correlation between predicted and actual opening weekend box office revenue for feature films, where predictions are based on query volume during the one-week period t weeks prior to a movie's release.

Next, we compare the search-only model to a baseline model that incorporates information on film production budgets, the number of opening screens, and revenue projections from the prediction market Hollywood Stock Exchange:

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

This baseline model predicts opening weekend revenue quite well, and in fact outperforms the search-only model. Specifically, the correlation between baseline predictions and actual outcomes is 0.94, compared to 0.85 for the search-based predictions made immediately prior to a film's release.[1]

Finally, we evaluate predictions from a model that combines search with the baseline features:

$$\log(\textbf{revenue}) = \beta_0 + \beta_1 \log(\textbf{budget}) + \beta_2 \log(\textbf{screens}) + \beta_3 \log(\textbf{hsx}) + \beta_4 \log(\textbf{search}) + \epsilon$$

The addition of search volume negligibly improves model fit over the baseline, measured either in terms of correlation (0.94) or MAE (0.66). Fits and model coefficients for the search, baseline, and combined models are summarized in Table 1. Predictions generated under these models are displayed in Figure 3.

In summary, relative to the informative baseline, we find that search has little marginal predictive value. Nevertheless, we also see that search activity alone is indicative of box-office revenue, and remains so even weeks in advance of a movie's release—an interesting phenomenon itself.

3.2 Video Games

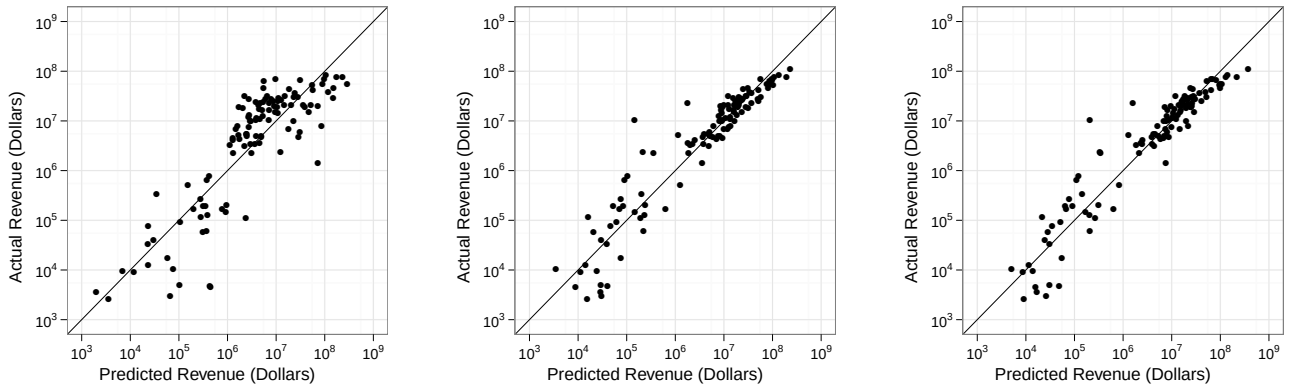
Data Description. We collected data on 106 video games released between September 2008 and September 2009. The task again is to predict the early commercial success of each game, quantified in this case as the total sales across all gaming platforms (e.g., XBox, Play Station) in the U.S. market in the first month following release. Sales data were obtained from VG Chartz (`vgchartz.com`), a leading video game sales tracking website. First month sales range from \$4K to \$2M, with a mean of \$2.9M and a median of \$1.1M. In addition to sales data, VG Chartz publishes average critic ratings, summarized on a scale from 1 to 10, from which we built a baseline prediction model. Ratings ranged from 3.1 to 9.5, with a mean of 7.5 and a median of 7.7. For the 38 games in our dataset that were sequels, we also considered the lifetime revenue of the game's predecessor, as reported by VG Chartz. We suspect that production budgets and initial distribution sizes would be highly informative features for predicting revenue; however, to the best of our knowledge, this information is not publicly available.

As in our analysis of movies, we measured user intent by examining the the first page of search results for each query. Specifically, we mapped queries to games by extracting game-specific identifiers from URLs for three leading gaming web sites: GameTrailers, GameSpot, and GamePro. Figure 4 plots query volume for *Tom Clancy's H.A.W.X.* during the one month period before and after the game's release. As with movies, one observes a sharp spike in search

[1]The baseline model is trained on HSX estimates reported one day prior to a movie's release. HSX estimates, like search, are updated daily, and it would be interesting to investigate the relative performance of search to HSX over time.

	Estimated Model					Fit	MAE
	Intercept	log(**search**)	log(**screens**)	log(**budget**)	log(**hsx**)		
Search	3.56 (0.66)	1.36 (0.07)				0.85	1.19
Budget+Screens+HSX	7.15 (1.53)		0.35 (0.05)	0.13 (0.10)	1.15 (0.11)	0.94	0.68
Budget+Screens+HSX+Search	6.82 (1.49)	0.27 (0.10)	0.33 (0.05)	0.06 (0.10)	0.94 (0.13)	0.94	0.66

Table 1: Summary of model coefficients and fits for movies, where the outcome variable is the logarithm of opening weekend box-office revenue. Fit is defined to be the correlation between predicted and actual (log-transformed) outcomes.



(a) A model that includes only search data.

(b) A baseline model that incorporates film production budgets, the number of opening screens, and forecasts from Hollywood Stock Exchange.

(c) A model that combines search and baseline features.

Figure 3: Predictions of opening weekend box office revenue from search-only, baseline, and combined models. Search volume is measured during the 7 days prior to a film's opening, and the baseline model includes revenue forecasts from Hollywood Stock Exchange available the day before a film's opening.

queries for the game around its release date, suggesting that our approach does capture user intent.

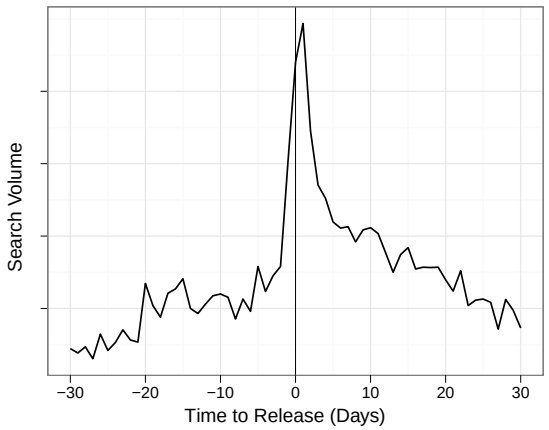


Figure 4: Search volume for the video game *Tom Clancy's H.A.W.X.* prior to and after its release.

Analysis. We fit simple linear models of first month sales against search volume, critic ratings, and predecessor sales (for sequels). With the exception of ratings, inputs and outputs were log-transformed for the analysis. As before, we consider correlation and MAE between predicted and actual (log-transformed) outcomes, where fitness is computed via leave-one-out cross-validation. A summary of model coefficients and fits is presented in Table 2, and model predictions are displayed in Figure 6.

We investigate the performance of search alone by estimating and evaluating the model:

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

Figure 5 shows that search has good predictive quality even several weeks before a game's release. For example, one month before a game's release, we find a correlation of 0.7 between predicted and actual outcomes.

We next compare the search-only model to a baseline model that incorporates average critic ratings:

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

This model does not, in fact, perform very well, exhibiting a correlation between predicted and actual outcomes of 0.43. On the one hand, since critics are not expressly trying to forecast sales performance, this result is not surprising. On the other hand, ratings can impact early sales, and furthermore, there do not appear to be many other publicly available features from which to construct a general base-

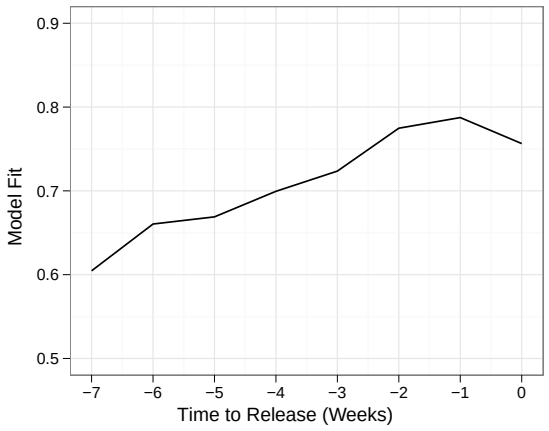


Figure 5: Correlation between predicted and actual video game sales, where predictions are based on query volume during the one-week period t weeks prior to a game's release.

line model for video games.[2]

Ratings add modest predictive value to the search-only model: The combined model

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

has correlation 0.80 between predicted and actual values—compared to 0.76 for the search only model—and MAE improves from 0.73 to 0.70.

We note that the relatively impressive performance of search over the baseline model is to a large extent an artifact of the particularly weak available baseline. We illustrate this point by analyzing the subset of games that are sequels, where baseline predictions are generated by considering the lifetime revenue of the game's predecessor:

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

For sequels, the correlation between predicted and actual values is 0.81, while the search-only model has a correlation of 0.68. The combined model

$$\log(\textbf{revenue}) = \beta_0 + \beta_1 \log(\textbf{predecessor}) + \beta_2 \textbf{rating} + \beta_3 \log(\textbf{search}) + \epsilon$$

offers negligible improvement over the baseline: The correlation between predicted and actual values rises from 0.81 to 0.82, and the MAE improves from 0.64 to 0.61.

While the baseline model of looking at predecessor performance applies only to sequels, it suggest that with additional baseline features (e.g., production budgets) one could obtain an improved generic model independent of search data. We also note that though sequels constitute just a third of our dataset, these include many of the most highly anticipated and top-selling games. Therefore, we are hesitant to conclude that search volume adds information of material value in predicting video game sales. However, in the absence of baseline information, it is interesting that search alone is a leading indicator of a game's commercial success.

[2]Part of the reason why ratings perform poorly is that they do not show much variation, with over half of the games in our dataset receiving a rating between 7 and 9. Moreover, ratings for highly anticipated franchise games tend to be even more tightly clustered, making it difficult for the baseline model to extract meaningful information.

3.3 Music

Data Description. We analyze the 307 songs that appeared on the Billboard Hot 100 list between March and September of 2009 with the goal of predicting song rank for future weeks. The Hot 100 chart is issued weekly by the magazine Billboard, and lists the 100 most popular singles in the U.S. based on airtime and sales. The data were acquired through the Billboard Developer API (`developer.billboard.com`) and include artist, song title, rank, and chart release date. We note that there is a substantial reporting delay in the Billboard charts, as the currently available chart provides information collected over a week ago; in our analysis below, we are careful to assume access only to information publicly available at the time of prediction.

Weekly search counts for queries were obtained from Yahoo! Music (`music.yahoo.com`) query logs for the U.S. market. For a given week we aggregated query volume for all songs appearing on the currently available Billboard Hot 100 chart.[3] In contrast to our analysis of movies and video games, where user intent is measured by examining the links that appear in search results, intent for music is quantified by tallying queries 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 Paisely) for which intent in general web search is harder to discern. Figure 7 provides an example time series of Billboard and search rank for a single song over a six month period.

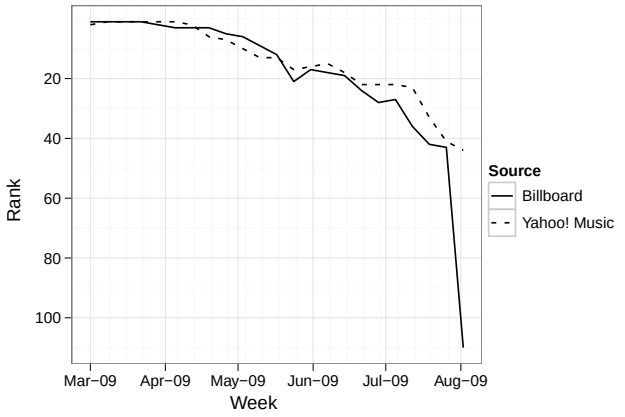


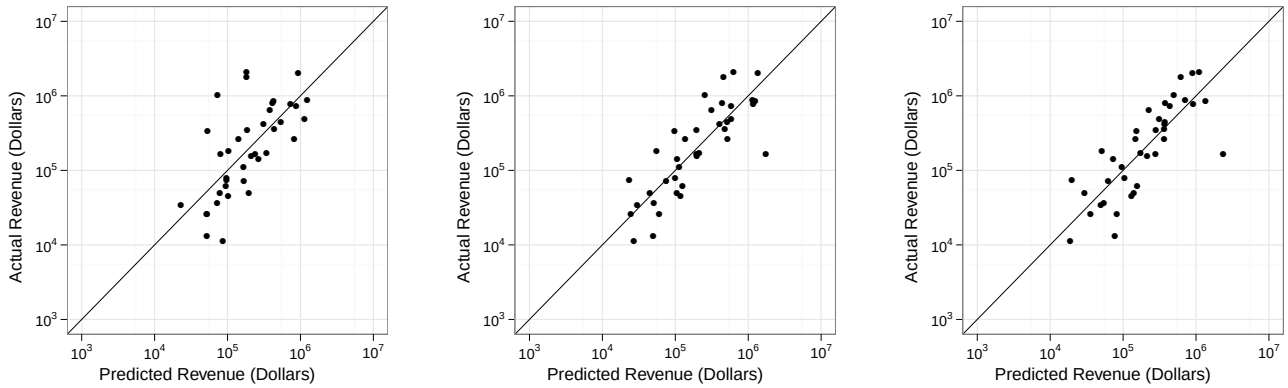
Figure 7: Search and Billboard rank for the song "Right Round" by Flo Rida.

Analysis. In keeping with our objective of prediction rather than faster reporting, we investigate the extent to which currently available search and Billboard data can be used to forecast future song performance. We consider three models of future song rank: a baseline that ignores search and uses only Billboard data, a model which uses only search

[3]In so doing, we implicitly abstain from predicting when a song first enters the Billboard charts, although one could proceed with the same analysis using external sources to obtain new song titles by release date.

	Estimated Model				Fit	MAE
	Intercept	log(**search**)	**rating**	log(**predecessor**)		
Search	8.28 (0.29)	0.55 (0.05)			0.76	0.73
Ratings	7.38 (0.82)		0.57 (0.11)		0.43	1.11
Ratings+Search	6.06 (0.55)	0.50 (0.04)	0.34 (0.07)		0.80	0.70
Search	8.48 (0.63)	0.55 (0.09)			0.68	0.80
Predecessor+Ratings	11.12 (0.91)		0.17 (0.12)	0.83 (0.10)	0.80	0.64
Predecessor+Ratings+Search	9.28 (1.12)	0.23 (0.09)	0.20 (0.11)	0.61 (0.13)	0.83	0.60

Table 2: Summary of model coefficients and fits for video games, where the outcome variable is the logarithm of first-month sales revenue. Fit is defined to be the correlation between predicted and actual (log-transformed) outcomes. The top three models predict revenue for all video games in our dataset, while the bottom three are restricted to sequels.



(a) Predicted revenue based on search volume in the week prior to game release.

(b) Baseline predictions based on total revenue for a game's predecessor and critic ratings.

(c) Predictions from a combined search and baseline model.

Figure 6: Predicted vs. actual revenue for first-month sales of video game sequels.

data, and a combined model that includes both Billboard and search data. In the two latter models, where search data were used, songs were rank-ordered by weekly query volume to create a search rank for the current and previous weeks.

In all three models we make predictions for the next week, denoted $t+1$, given information from the current and previous weeks, denoted t and $t-1$, respectively.[4] For each week in the test set, we train on the previous three months and evaluate the following week's prediction. This results in 13 weeks of test data from June through August. The coefficients, test correlations, and mean absolute error (MAE) in predicted rank for all three models are reported in Table 3.

As a baseline, we used a simple autoregressive model that predicts next week's Billboard rank as a linear function of the currently available Billboard rank:

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

This baseline model performs relatively well, with a correlation of 0.70 between predicted and actual rank across all songs and weeks in the test set. From Figure 9(b) we see that many of the errors made by the baseline model are from high-ranking songs that quickly fall off the charts, for which the predicted rank is much higher than the actual rank.

We compare this to a search-only model, with Billboard rank omitted as a predictor:

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

This model yields a reasonable test set correlation of 0.56, which only slightly deteriorates as predictions are made further in advance of the outcome in question—see Figure 8. However, the performance is significantly lower than that of the search-only models for movies (0.85) and video games (0.76). From the coefficients in Table 3, we see that this model takes a weighted difference between the current and previous week's search rank. Though such a model is useful for predicting a relative change in Billboard rank, it is not as well suited to predicting absolute rank.

Finally, we find that the combined model

$$\mathbf{billboard}_{t+1} = \beta_0 + \beta_1 \mathbf{search}_t + \beta_2 \mathbf{search}_{t-1} + \beta_3 \mathbf{billboard}_{t-1} + \epsilon$$

that includes both currently available Billboard and search ranks performs significantly better than baseline or search-only models, with a correlation of 0.87. The combined model

[4]For songs on the current Billboard chart that were not on the following week's chart, the following week's Billboard rank is coded as 110, as the average week-to-week change in rank is 10.

admits a relatively simple interpretation: The currently available Billboard rank anchors the prediction, and the change in search rank indicates movement relative to this baseline. This eliminates many of the large errors between predicted and actual values, as shown in Figure 9(c).

In interpreting the performance of the combined model (0.87) over the baseline (0.70), we emphasize that due to delayed reporting by Billboard, the baseline model cannot incorporate chart ranking information for the current week (i.e., $\mathbf{billboard}_t$). If this information were available at the time of prediction, however, an autoregressive baseline would perform substantially better, and the boost from search data would be greatly reduced. We quantify this with an *ex post facto* analysis which removes the lag in Billboard reporting; the result is a baseline with test correlation of 0.91, while the combined model achieves a marginal boost to 0.92.

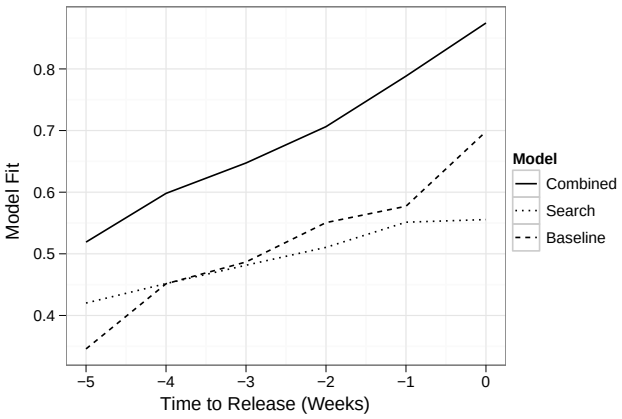


Figure 8: Correlation between predicted and actual Billboard rank for songs, where predictions are based on query volume during the one-week period t weeks prior to the predicted outcome.

As we have seen in the movie and video game domains, the simple baseline model shows reasonably good performance—a song's current Billboard rank is highly predictive of its future rank. In the absence of this baseline, current search activity for a song is still somewhat predictive of its future Billboard rank. Performance improves substantially by combining the baseline and search models, where Billboard ranks anchor predictions and relative changes in search ranks indicate chart movement.

4. RELATIVE PERFORMANCE

Across the three empirical cases considered above, we observe a consistent pattern. First, when search counts are considered in isolation, they are predictive of human activities—like attending movies, downloading music, or purchasing video games—that will take place days, or in some cases even weeks, in the future. Second, however, they are generally less predictive than simple baseline models that use other publicly available information. And, third, when used in combination with such baselines, search counts generally contribute a small boost to model performance.

Although our findings regarding the performance of search-based predictions appear less impressive when considered in the context of the appropriate baseline model, we note that they are generally consistent with other recent findings in the search prediction literature, in particular those of Choi and Varian [3, 2] who also find that search counts provide at most a small boost over simple models in predicting economic activity. More generally still, this pattern is also consistent with a number of previous studies of forecasting [11, 10, 12], which in general find that simple models perform approximately as well as sophisticated alternatives. Finally, as we illustrate next, the same pattern applies also to the example of influenza caseload data, which is arguably the best-known example of search queries correlating with real-world outcomes.

Data Description. As we noted in the introduction, previous studies correlating flu-related terms to influenza caseloads [7, 13, 8] have not evaluated performance relative to that of any baseline model; thus we now compare search-based estimates of influenza as reported on Google Flu Trends [8] to a simple autoregressive model. Ground truth incidence levels were obtained from the Centers for Disease Control (CDC). As public reports of flu caseloads by the CDC are typically delayed 1-2 weeks, the primary aim of search-based flu estimation is real-time monitoring, as opposed to strictly predicting future activity. In particular, Flu Trends estimates for a given week rely on query volume during that same week; thus, 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.

Analysis. Given the reporting delay of flu caseloads, our baseline autoregressive model relies only on data for periods at least two weeks prior to the current week:

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

At the end of week t—when Flu Trends releases an estimate for week t—the autoregressive model uses the most recent report of flu incidence, which details ground truth data from week $t-2$.

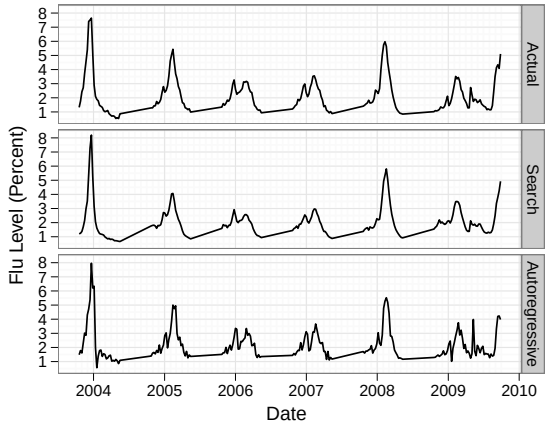
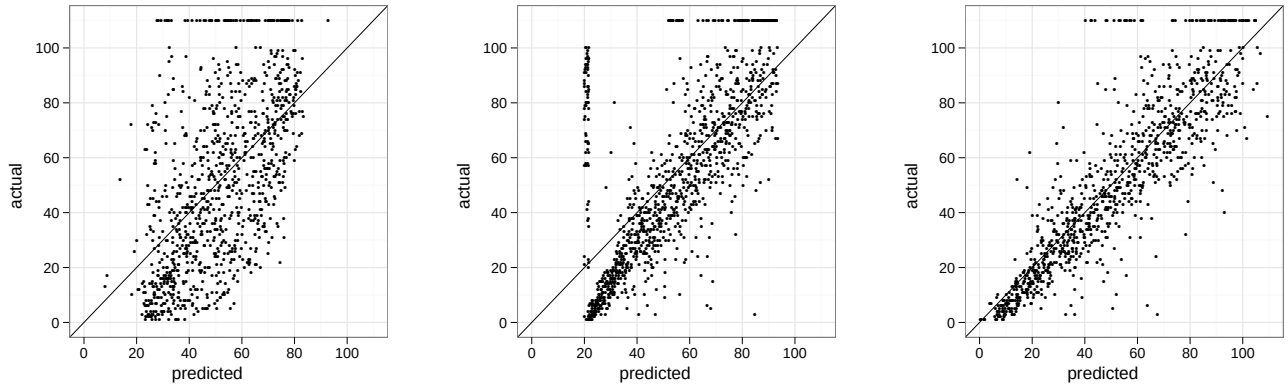


Figure 10: Actual and estimated flu levels in the U.S., where flu level is the percentage of physician visits that involve patients with influenza-like ilnesses. Search-based estimates are from Google Flu Trends.

As shown in Figure 10, both the search-based estimates and the baseline model perform quite well at monitoring flu

	Estimated Model				Fit	MAE
	Intercept	$\mathbf{search}_t$	$\mathbf{search}_{t-1}$	$\mathbf{billboard}_{t-1}$		
Search	22.90 (1.13)	1.24 (0.11)	-0.66 (0.11)		0.56	20.62
Billboard	21.16 (0.86)			0.71 (0.02)	0.70	17.05
Search+Billboard	5.67 (0.73)	1.11 (0.06)	-1.09 (0.06)	0.93 (0.01)	0.87	10.48

Table 3: Summary of model coefficients and fits for music, where the outcome variable is song rank on the Billboard Hot 100 chart. Fit is defined to be the correlation between predicted and actual outcomes.



(a) Future Billboard rank is based on search volume during the current and previous weeks.

(b) A baseline model for future Billboard rank, where predictions are based on the current week's Billboard rank.

(c) Predicted Billboard rank is based on the current week's Billboard rank and search volume for the current and previous weeks.

Figure 9: Predicted vs. actual Billboard rank for music data.

incidence. Specifically, the search estimates yield a correlation of 0.94 between actual and predicted outcomes, while the baseline model has a correlation of 0.86. Thus, though the improvement of search over the baseline is potentially consequential to public health surveillance, these results illustrate that weekly variability in flu incidence is estimable to a large extent through very simple means. We further note that if one assumes flu reports are delayed by only one week instead of two, then the corresponding autoregressive model has a fit of 0.95, in fact outperforming the search-based estimates.

5. CONCLUSION

Returning to the motivating question of this paper—What can search predict?—we have shown that web search is a (perhaps surprisingly) accurate leading indicator of consumer activities, comparable in quality to other public reports. Looking across a range of recent findings, a broadly consistent story emerges: In the absence of a readily available baseline model, search counts can generate useful predictions; however, they do not generate dramatically better predictions than those from other, simple means. Figure 11 summarizes these findings, showing the predictive fit of search, baseline, and combined models across all the domains we study: movies, video games, music, and flu.

Whether search counts are useful for predicting cultural events is therefore a matter of circumstance and necessity: On the one hand, the performance of the combined models is often not qualitatively different from the baseline; but on the other hand, it may well be the case that in certain situations, such baseline estimates are difficult to generate, or that for some applications, even small gains in performance are valuable. Put another way, the main advantage of web search may have less to do with its superiority over other data sources than with its generality, low cost, and real-time nature.

Finally, we conclude by noting that not all activities of interest lend themselves equally to search-based predictions, either because they differ in how well they are reflected in search behavior, or because of varying difficulty in identifying user intent. Movies, for example, generally generate high search volume, reflecting the large number of people who are interested in finding the location and timing of even a moderately successful feature film, and queries for movies are relatively easy to identify. Songs by contrast, elicit fewer searches, presumably in part because there are many more individual songs than movies at any point in time. Song names, moreover, are frequently indistinguishable from ordinary english phrases; thus the intent to search for a song is more difficult to infer. For both these reasons, it is not surprising that we find correlations between search counts and movie revenues to be considerably higher than for song rankings. More generally, because search is observed to correlate well with activities in one domain does not necessarily imply that it works well in all domains, or that the methods for extracting the relevant data will be broadly effective. Although the likely domain dependency of search-based prediction does raise some additional challenges, we hope that

these and other issues will be addressed satisfactorily in future work.

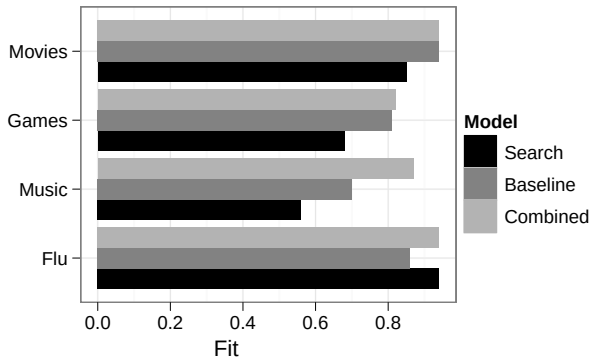


Figure 11: The correlation between predicted and actual outcomes for movies, video game sequels, music, and flu.

Acknowledgments

We thank Jitendra Waral for pointing us to VG Chartz, and we thank Vanessa Colella and Nick Weir for encouraging conversations.

6. REFERENCES

[1] N. Askitas and K. F. Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, 2009.

[2] H. Choi and H. Varian. Predicting initial claims for unemployment benefits. Available at `http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf`, July 2009.

[3] H. Choi and H. Varian. Predicting the present with google trends. Available at `http://research.google.com/archive/papers/initialclaimsUS.pdf`, April 2009.

[4] C. Cooper, K. Mallon, S. Leadbetter, L. Pollack, and L. Peipins. Cancer internet search activity on a major search engine, united states 2001-2003. *J Med Internet Res*, 7, 2005.

[5] C. D. Corley, A. R. Mikler, K. P. Singh, and D. J. Cook. Monitoring influenza trends through mining social media. In *Proceedings of the 2009 International Conference on Bioinformatics and Computational Biology (BIOCOMP09)*, 2009.

[6] M. Ettredge, J. Gerdes, and G. Karuga. Using web-based search data to predict macroeconomic statistics. *Commun. ACM*, 48(11):87–92, 2005.

[7] G. Eysenbach. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annu Symp Proc*, pages 244–248, 2006.

[8] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, November 2008.

[9] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4, 2009.

[10] S. Makridakis and M. Hibon. The m3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16:451–476, 2000.

[11] S. Makridakis, M. Hibon, and C. Moser. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society: Series A*, 142:97–145, 1979.

[12] S. Makridakis, R. M. Hogarth, and A. Gaba. Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, 2009.

[13] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47:1443–1448, 2008.

[14] B. J. Wilson K. Early detection of disease outbreaks using the internet. *CMAJ*, 2009.

[15] J. Wolfers and E. Zitzewitz. Prediction markets. *The Journal of Economic Perspectives*, 18:107–126, 2004.