The Future of Prediction:

How Google Searches Foreshadow Housing Prices and Sales

Lynn Wu
MIT Sloan School of Management
50 Memorial Drive, E53-314
Cambridge, MA 02142
wulynn@wharton.upenn.edu

Erik Brynjolfsson
MIT Sloan School of Management
50 Memorial Drive, E53-313
Cambridge, MA 02142
erikb@mit.edu

This Draft: April 15, 2013

Comments Welcome

Abstract

Most data sources used in economics, whether from the government or businesses, are typically available only after a substantial lag, at a high level of aggregation, and for variables that were specified and collected in advance. This hampers the effectiveness of real-time predictions. We demonstrate how data from search engines like Google provide a highly accurate but simple way to predict future business activities. Applying our methodology to predict housing market trends, we find that a housing search index is strongly predictive of the future housing market sales and prices. The use of search data produces out-of-sample predictions with a smaller mean absolute error than the baseline model that uses conventional data but does not include any search data. The improvements in predictions using search terms is 7.1% better over the baseline for future home sales and 4.6% better for future HPI. Furthermore, we find that our simple model of using search frequencies can out-predict experts from the National Association of Realtors by 23.6% for predicting future US home sales. We also demonstrate how these data can be used in other markets, such as laptop sales. In the near future, this type of "nanoeconomic" data can transform prediction in numerous markets, and thus business and consumer decision-making.

Keywords: Online Search, Prediction, Housing Trends

"It's difficult to make predictions, especially about the future"
-- Attributed to Neils Bohr

Introduction

Traditional economic and business forecasting has relied on statistics gathered by government agencies, annual reports and financial statements. Invariably, these are published after significant delay and are aggregated into a relatively small number of pre-specified categories. This limits their usefulness for predictions, especially novel predictions. However, due to the widespread adoption of search engines and related information technologies, it is increasingly possible to obtain highly disaggregated data on literally hundreds of billions[1] of economic decisions almost the instant that they are made. Now, query technology has made it possible to obtain such information at nearly zero cost, virtually instantaneously and at fine-grained level of disaggregation. Each time a consumer or business decides to search for a product via the Internet, valuable information is revealed about that individual's intentions to make an economic transaction (Moe & Faber, 2004). In turn, knowledge of these intentions can be used to predict demand, supply or both. This revolution in information and information technology is well underway and it portends a concomitant revolution in our ability to make business predictions and ultimately a sea change in business decision-making. This new use of information technology is not a mere difference in degree, but a fundamental transformation of what is known about the present and what can be known about the future.

Assisting with predictions has always been a central contribution of social science research. In the past several decades, much of social science research has focused on ever more complex mathematical models, for many types of important business and economic predictions. However, the latest recession has shown that none of the theoretical models was intelligent enough to foresee the biggest economic downturn in our recent history (Krugman, 2009). Perhaps, instead of honing techniques to extract information out of noisy and error-prone data, social science research should focus on inventing tools to observe phenomenon at a higher resolution (Simon, 1984). Search engine technology has precisely delivered such a tool. By effectively aggregating consumers' digital traces and improving data quality by several orders of magnitude, information technology has created a transformation on how we solve the problem of predicting the future. With the observation of billions of consumers and business intentions

[1] Americans performed 14.3 billion Internet searches in March, 2009, which is an annualized rate of over 170 billion searches per year. Worldwide searchers group by 41% between 2008 and 2009.

as revealed by online search, we can significantly improve the accuracy of predictions about future economic activities.

In this paper, we demonstrate how data on Internet queries could be used to make reliable predictions about both prices and quantities literally months before they actually change in the marketplace. We use the housing market as our case example but our techniques can be applied to almost any market where Internet search is non-trivial, which is to say, an increasingly large share of the economy. What's more, by identifying correlations with prices and quantities we can make inferences about changes in the underlying supply and demand. Our techniques can be focused on particular regions or specific cities, or the nation as a whole, and can look at broad or narrow product categories. Search not only precedes purchase decisions, but in many cases is a more "honest signal" (Pentland, 2008) of actual interests and preferences since there is no bargaining, gaming or strategic signaling involved, in contrast to many market-based transactions. As a result, these digital traces left by consumers can be compiled to reveal comprehensive pictures of the true underlying economic intentions and activities. Using aggregated query data collected from the Internet has the potential to make accurate predictions about areas as diverse as the eventual winners of standard wars, or the potential success of product introductions.

We started making housing market predictions since January of 2009 and as of September 2011, almost three years after we released our first set of real estate predictions, search queries continue to have a strong impact in forecasting real estate trends and outperform predictions from the National Association of Realtors. If search queries were simply an ephemeral phenomenon, we would expect their informational value to quickly be absorbed into the market. However, search queries continue to be relevant years after search queries are used for predictions. This shows the persistence of the economic value derived from search that is beyond the natural equilibrium consequences.

The Real Estate Market

We use the real estate market to demonstrate how online search can be used to reveal the present economic activities and predict future economic trends. Studying the real estate market is especially important at the awake of the recent burst of the real estate bubble that has triggered the current economic downturn in the US and the rest of the world. In turn, when the housing market becomes healthy again, the economy may be on the mend as well (New Work

Times Editorial, 2009). Economists, politicians and investors alike are pouring over government data released every month to assess the current housing market and predict its recovery and subsequently the revival of economic growth. However, government data are often released with a lag of months or more, rendering a delay in assessing the current economic conditions. We propose a different way to reveal the present and predict the future housing trends via the frequency of online search terms. Analyzing consumers' interests as revealed by their online behaviors, we are able to uncover trends before they appear in published data.

The Internet is a valuable research tool and can provide critical information to make purchase decisions (Horrigan, 2008). As the Web becomes ubiquitous, more shoppers are using the Internet to gather information and narrow down the number of selections, especially for products that require a high level of financial commitment, such as buying a home. According to the 2012 Profile of Home Buyers and Sellers by National Association of Realtors (NAR), 90% of home buyers used the Internet to search for a home in 2012 (NAR, 2012). Similarly, a report, written by California Association of Realtors in 2008, shows that 63% of homebuyers find their real estate agent using a search engine (Appleton-Young, 2008). To explore the link between search and actual sales, we analyze billions of individual searches from five years of data in the Google Web Search portal[2] to predict housing sales and housing prices. Using these fine-grained data on individual consumer behaviors, we built a comprehensive model to predict housing market trends.

We found evidence that queries submitted to Google's Search Engine are correlated with both the volume of housing sales as well as a house price index—specifically the Case-Shiller Index. The Case-Shiller index is a predominant housing index and is widely used in most government reports. We find that search frequencies can reveal the current housing trends but it is especially well suited for predicting the future housing sales. Specifically, we find that a one-percentage point increase in search frequency about real estate agents is associated with selling an additional 3,520 future quarterly housing sales in the average US state. We also compared our predictions with the prediction released by the National Association of Realtors (NAR) and our simple linear prediction model using search frequencies outperforms NAR's predictions by 23%.

[2] http://www.google.com/insights/search/#

Similarly, we also examine the relationship between housing price and housing related searches online. Using house price index (HPI) from Federal Housing Finance Agency,[3] we find a positive relationship between the housing related online queries and the future house price index, though the predictive power is not as strong as it is for home sales. Perhaps, predicting HPI is considerable harder than sales volume because it is unclear how the search volume can affect HPI. If the search volume happens to move the demand curve, HPI will go up. On the other hand, if the search volume comes from the supply side, HPI can go down. Thus, the search volume could either increase or decrease HPI depending on whether search terms are related to the demand or the supply side. Aggregated search indices on general real estate categories may be well suited to predict sales but not for differentiating between a supply side change and a demand side change. Thus, a less aggregated and more fine-grained search categories could be created to differentiating the shifts on the demand side from the supply side.

We also find evidence that the total volume of houses sold is correlated with consumers' intention to purchase home appliances. We use the search frequency of home appliances to approximate their consumers' interests (Moe and Fader, 2004). We find that every thousand houses sold are correlated with a 1.23 percentage point increase in the frequency of search terms that are related to home appliances. This highlights the linkages between home sales and other parts of the economy that complement home sales.

Literature Review

In the past decades, much of the social science research has focused on refining increasingly complex mathematical models to predict social and economic trends. However, the alternative of collecting high quality data at a much finer-grained levels has mostly eluded in social science research.

Today, advances in information technologies, such as the Internet search technologies, e-mail, smart sociometric badges, social media, offer remarkable detailed records of human behaviors. Recently, researchers have started to take advantage of real-time data collected from these new technologies. For example, deploying sociometric badges to measure moment-to-moment interactions among a group of IT workers, Wu et al. (2008) has uncovered new social network dynamics that are only possible by accessing accurate data at micro-level. Lazer et al (2009)

[3] http://www.fhfa.gov/

provided various examples of how high quality data produced by novel technologies are transforming the landscape of social network research. Similarly, firms have also leveraged the massive amounts of data collected online to make predictions, such as consumer preferences, supplies and demands for various goods as well as basic operational parameters such as inventory level and turnover rate. The ability to collect and efficiently analyze the enormous amount of data made available by information technology has enabled firms, such as Amazon, Harrah's and Capital One, to hone their business strategies and to achieve tremendous gain in profitability and market shares (Davenport, 2006).

Our work follows a similar stream in demonstrating the power of using fine-grained data to predict underlying social and economic trends. Unlike previous research and businesses that have primarily used proprietary data, we leverage free and public available data from Google to accurately forecast economic trends. Research has shown that online behaviors can be used to reveal consumers' intention and predict purchase outcomes (e.g. Moe and Fader, 2004; Kuruzovich et al. 2008). We believe that we can rely on digital traces left by trillions of online search to reveal consumers' intentions and examine their power to predict underlying social and economic trends. Using such fine-grained data to study individual buying or selling decisions could be called nanoeconomics. Furthermore, it is unlikely that the effect of search in predictions will disappear once its informational value is absorbed into economic equilibria. We have been making predictions since 2008, and three years later, search frequencies continue to have an impact predicting real estate trends, and better than expert predictions from the National Association of Realtors.

We believe that we are only at the beginning of the data revolution. Newer and more fine-grained data are becoming available everyday from various search, social media and micro-blogging platforms. These data are made available instantaneously, allowing consumers and policy makers to tap on the pulse of economic activities as they are happening. However, their instantaneous nature can also do a lot of harm if not handled appropriately, as the recent Twitter hacks on the stock market have shown. Aggregating and integrating these nano-data in a meaningful fashion and teasing real economic data from the faulty ones are thus the key to make better predictions. Predicting instantaneous changes on the stock market may prove to be difficult due to quick changes on the market. However, predicting medium or longer-term trends, such as movements in the real estate market, could be easier because they are less prone to short-term manipulations, such as fake Twitter feeds that go viral quickly but die down shortly after they are revealed to be false .

Our methodologies are similar to a recent analysis on flu outbreaks using Google Flu Trends (Ginsberg et, al., 2009) and also parallel, but unpublished research by Choi and Varian (2009) where the authors also correlate housing trends in the US using search frequencies. While Choi and Varian (2009) mainly focus using search frequencies to reveal the current economic statistics, our work attempts to predict the *future* economic trends, such as forecasting price and quantity of houses sold in the future. At least in the real estate setting, we show that using search is more powerful for predicting the future than for predicting the present. Furthermore, our work also use more fine-grained data at the state level instead of at the level of the whole nation to provide a more nuanced prediction of real estate market which often varies greatly depending on geographical locations. Our future work will expand the analysis to the metropolitan statistical areas.

Economics of Real Estate

Our work also contributes to the literature of real estate economics. There are two types of forecasting methodologies for predicting real estate market trends. The first is the technically analysis that is used to predict stock market trends. The main assumption for this type of analysis is that the key statistical regularities for the underlying housing price do not change. The trending behaviors are therefore more likely to exhibit long-term reversion to the mean but with short-term momentum (e.g. Case & Shiller, 1989). Glaeser and Gyourko (2007) found evidence of long-term reversion in housing price. They found that, ceteris paribus, when regional prices go up by an extra dollar over one five-year period, the regional price on average would drop by 32 cents over the next five years. The second approach to predicting housing market trends is to use fundamental economic analysis. Housing price should depend heavily on the cost of construction, the interest rate to finance the housing purchase, the regional income as well as the January temperature (Glaeser, 2009). In principle, this suggests that regions with steady building costs and relatively stable income level should have a steady housing price. However, these economic variables do not seem to fully capture housing price trends. For instance, in Dallas, an example of a region with steady fundamentals, the housing price has been increasing despite the predictions of fundamental analysis.

Some dynamic housing demand models try to incorporate both approaches to predict housing trends (Glaeser and Gyourko 2007, Han 2009). Using dynamic rational expectation to model housing price, Glaeser and Gyourko (2007)

detects mean-reverting mechanism but they cannot explain serial correlation or price changes in most volatile markets. Glaeser (2009) suggests this may reflect sentiment or even "irrational exuberance" in some housing markets, generating a bigger boom and bust cycle than what are predicted by the model (Glaeser 2009).

With the ability to gather billions of search queries over time, Google Trends is essentially aggregating all the honest signals of decision-makers' intentions to capture the overall level of "sentiment". This provides unprecedented opportunities to improve predictions in housing markets. Using very simple regression models, we demonstrate that Google search frequencies can be used as a reliable predictor for the underlying housing market trends both in the present and in the future.

Data Sources

Google Search Data

We collected the volume of Internet search queries related to real estate from Google Trends, which provides weekly and monthly reports on query statistics for various industries. It allows users to obtain a query index pertaining to a specific phrase such as "Housing Price". Google Trends has also systematically captured online queries and categorized them into several predefined categories such as Computer & Electronics, Finance & Business and Real Estate. As Nielsen NetRatings has consistently placed Google to be the top search engine, which processed more than 66.7% of all the online queries in the world in December 2012 (comScore, 2012), the volume of queries submitted to Google has the potential to approximate people's interests over time. In fact, recent work using Google Search can accurately predicts flu outbreaks few days before it actually happens (Ginsberg et al, 2008). We believe that the search volumes can also be used to predict future economic indicators, such as real estate trends.

Google Trends provides a search index for the volume of queries based on geographic locations and time. The search index is a compilation of all Internet queries submitted to Google's search engine since 2004. The index for each query phrase is not the absolute number of queries submitted. Instead, it reports a query index measured by query share, which is calculated as the search volume for the query in a given geographical location divided by the

total number of queries in that region at a given point in time[4]. Thus, the index is always a number between 0 and 100. The reports on search indices are also much more fine-grained than most government reports. Typically, Google calculates the query index on a weekly or a monthly basis and can be disaggregated down to country, state/province and city levels around the world. For example, in the US, a query index can be calculated at the state level. A more detailed query index at the MSA level can also be computed by specifying the appropriate sub-regions within a state. Figure 1 shows the overall interest in the search category "Real Estate Agencies" using online searches in the US. From the graph, interests in housing peaked in 2005 and fell through 2009.

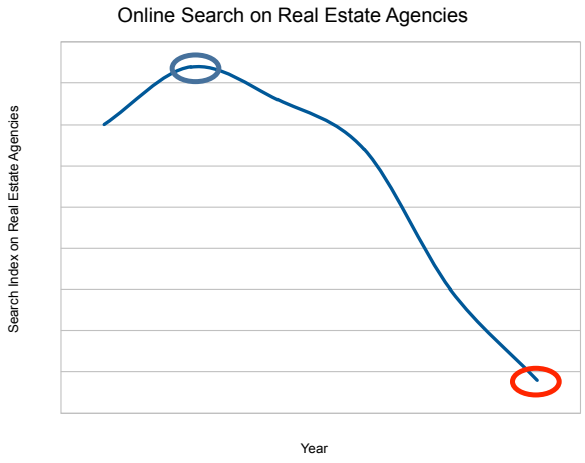


Figure 1: Search Index for "Real Estate Agencies. It is a normalized measure of search volume ranging from 0 to 100.

[4] http://www.google.com/support/insights//bin/answer.py?answer=87285

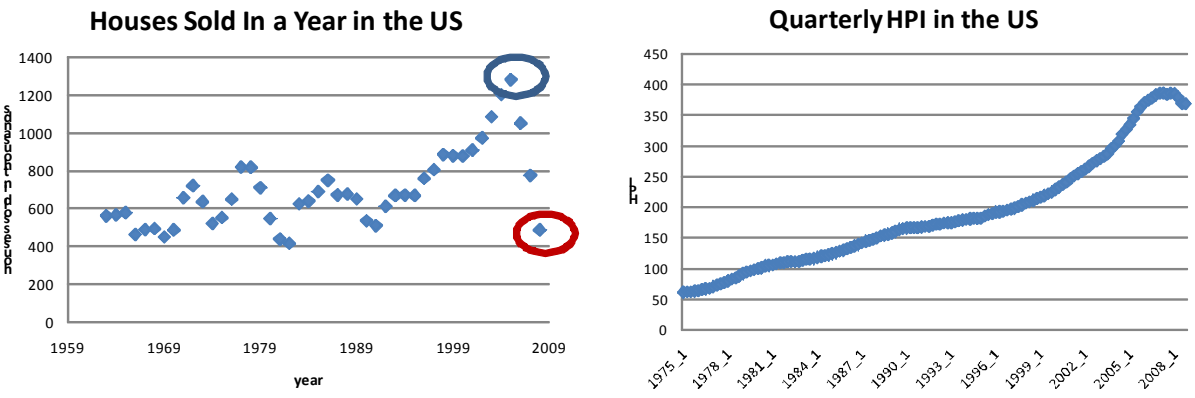


Figure 2: Housing and Prices of New House Sold in the US. (a) Number of New Houses Sold Annually. (b) Quarterly House Price Index.

Our analysis uses a predefined category in Google Trends, "Real Estate" to approximate the overall interest for housing. This category aggregates all online search queries that are related to real estate. We also collected more fine-grained search index for several subcategories, such as the "Real Estate Agencies" and "Real Estate Listings." [5] Furthermore, we compile our own bags of words related to various housing related transactions such as "housing sales", "home staging", or "home inspection." We hypothesize that these housing related search indices are correlated with the underlying conditions of the US housing market. To test this hypothesis, we gather housing market indicators, such as the volume of houses sold and the house price index in each US state, all from publicly available sources.

Housing Market Indicators

We collect the volume of housing sales from National Association of Realtors (http://www.realtor.org/research) for all 50 states in the US and the District of Columbia from the 1st quarter of 2006 to the 3rd quarter of 2011. We did not collect more data after 2011 mostly because we would like to compare our predictions with expert predictions from the National Association of Realtors (NAR). NAR has made their predictions public since 2005 but it has stopped publishing them after the 3rd quarter of 2011. We also obtain the house price index for the same period at

[5] We explored various pre-defined categories on Google Trends: "Apartments & Residential Rentals", "Commercial & Investment Real Estate", "Property Management", "Property Inspection & Appraisals", "Property Development", "Real Estate Agencies", "Real Estate Listings", "Timeshares & Vacation Properties."

the Office of Federal Housing Enterprise Oversight (http://www.ofheo.gov/), where housing prices for nine Census Bureau divisions are collected. The Office of Federal Housing Enterprise Oversight calculated the HPI for each state in the US on a quarterly basis since 1975[6]. Detailed calculations of the HPI can be found at http://www.fhfa.gov/.

As shown in Figure 2(a), the number of houses sold in the US peaked at around 2005 and then declined precipitously soon after, reaching a historically low at the beginning of 2009. The HPI also increased gradually and reached a peak in 2007, two years after the housing sales peak (Figure 2b), and began to fall shortly after. Comparing housing market indicators (Figure 2) to their associated online search indices (Figure 1) shows that they appear to be closely correlated. As shown in Figure 1, housing related search peaked at 2005 and gradually declined to its lowest point in early 2009, mirroring the volume of houses sold in Figure 2(a). This provides some evidence that the search indices are correlated with underlying housing trends.

Empirical Methods

First, we show that search indices are highly correlated with the underlying housing trends. We use a simple seasonal autoregressive (AR) model to estimate the relationship between search indices and housing market indicators— the volume of housing sales, and the house price index (HPI). A single class of explanatory variable is studied: search indices for housing related queries for each state in the US. We first estimate the baseline model to predict the current housing sales using only the past home sales and the past HPI. Then, we add the search indices to see if they improve predicting the contemporaneous home sales.

$$HomeSales_{it} = \alpha + \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{it,-1} + \beta_3 Population + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (1)$$

$$HomeSales_{it} = \alpha + \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{i,t-1} + \beta_3 SearchFreq_{it} + \beta_4 SearchFreq_{i,t-1} + \beta_5 Population + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (2)$$

We then examine whether the housing related search indices could forecast future housing sales. We use the past housing statistics to predict the future housing trends, because the present housing sales and indices are not available. Essentially, we are using a two-period lag to predict the future as opposed to a one-period lag to predict

[6] http://www.fhfa.gov/Default.aspx?Page=81

the present. While the government statistics are released with a lag, search frequencies on housing related inquiries are available in real-time and instantaneously down to the daily level. We can thus use both the present and the past search indices to predict future housing sales. Specifically, we use both one-period and two-period lags in the model because they are the most relevant for predictions. Higher order lags fail to have much predictive power. Presumably housing searches 9 months or one year earlier are too late to predict the present and the future housing trends as most of these searches might have already resulted in purchase decisions.

$$HomeSales_{it+1} = \alpha + \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{i,t-1} + \beta_3 SearchFreq_{it} + \beta_4 SearchFreq_{i,t-1} + \beta_4 SearchFreq_{i,t-2} + \beta_5 Population + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (3)$$

Similarly, we use the same approach to predict the current and the future HPI. In the baseline model, we only use the past HPI and past housing sales to predict the current HPI. We then incorporate the current and past search indices into the baseline model to predict the present HPI.

$$HPI_{it} = \alpha + \beta_1 HPI_{i,t-1} + \beta_2 HomeSales_{i,t-1} + \beta_3 Population\ \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (4)$$

$$HPI_{it} = \alpha + \beta_1 HPI_{i,t-1} + \beta_2 HomeSales_{i,t-1}\ \beta_3 SearchFreq_{it} + \beta_4 SearchFreq_{i,t-1} + \beta_5 Population + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (5)$$

Lastly, we predict the future HPI using the past HPI as well as the present and past search indices. In addition to exploring various lags, we also explored nonlinear functions of the search indices to see if they improve model fit and predictions.

$$HPI_{it+1} = \alpha + \beta_1 HomeSales_{i,t-1} + HPI_{i,t-1} + \beta_2 SearchFreq_{it} + \beta_3 SearchFreq_{i,t-1} + \beta_5 Population + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it} \quad (6)$$

For all the models above, we apply state- and region-level dummies in order to control for any time-invariant influences, such as the demographics of a state/region, and any statewide/region-wide policies that may affect real estate purchase decisions. We then train these models using data between the 1st quarter of 2006 and the 4th quarter of 2008 to find a set of search indices that best predict the present and future housing indicators. We then use these indices and their associated estimates to predict housing trends from the 1st quarter of 2009 to the 3rd quarter of 2011. For each prediction, we calculate the mean absolute error (MAE), defined below, to compare how accurate

our model using search indices compared to the baseline model as well as from predictions released by the National Association of Realtors. The mean absolute error is simply the deviation away from the actual value.

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \qquad (7)$$

In addition to housing predictions, we also examine whether housing related search queries can also spur future economic activities in complementary industries. For example, if consumers' intentions can be revealed through online search (Moe & Fader, 2004), we may also expect a surge in Internet queries about home appliances after observing a rise in home sales. Since new homeowners may plan to purchase appliances to furnish their property, tracking their online search behavior allow us to detect their intention to purchase home appliances. Accordingly, we correlate housing sales with the search index for home appliances. If search index for home appliance can translate into actual purchases, we would expect a rise in search frequencies for home appliances, spurred from home sales, to indicate a rise in their future demands as well.

$$HomeApplianceSearch_{it} = \alpha + \beta_1 HouseSold_{it} + \beta_2 HouseSolds_{i,t-1} + \varepsilon_{it} \qquad (8)$$

Empirical Results

First, we compare predictions between the baseline model and the model that uses search indices. We used the model to predict the present home sales and HPI as well as the future home sales and HPI in the next quarter. Although our model can be used to predict even more fine-grained forecasts, such as monthly or even weekly housing trends, we choose to forecast at the quarterly level because government statistics only release state-level housing sales and HPI every quarter. In order to calculate how accurate our predictions are, we aggregated the weekly search data into quarterly data. Furthermore, we also compare our predictions with the predictions released by the National Association of Realtors (NAR) that also forecasts quarterly housing sales. Unfortunately, NAR does not predict future HPI and thus we cannot compare our model with NAR's when predicting the future HPI.

Predicting Home Sales Using Online Search

Using search indices from Google, we find a positive relationship between housing sales and housing related search indices (Table 1). All models in Table 1 are based on a seasonal autoregressive (AR) model, which assumes that the sales in the future are related to sales in the past. We see a broad support for the AR model as the lagged sales are strongly correlated with the contemporary sales. We also applied a state-level fixed-effect specification to eliminate influence from any time-invariant factors and use seasonality dummies to control for time-specific changes. In addition, we also included the state population and region dummies to improve the fit of the model. To capture online interests for purchasing real estate properties, we use a search index of a predefined category in Google Trends – "Real Estate Listing" – that contains all queries pertaining to real estate listings and advertisements. We also use the "Real Estate Agencies" category to approximate home buying activities. We assume people who are looking for real estate agents and real estate listings online are more likely participate in a real estate transaction than those who search for other related queries such as property management.

First, we estimate the baseline model to predict the present home sales using only the past home sales and the past HPI. As shown in the baseline AR(1) model (Model 0), the past housing price and sales are highly correlated with the current home sales. We then examine various search indices related to the real estate market[7] and find two categories—"Real Estate Agencies" and "Real Estate Listings—to best predict the contemporaneous sales. Overall, we find that the contemporaneous search indices for "Real Estate Agencies" and "Real Estate Listings" are statistically significantly correlated with the present home sales. As shown in Model 1, a one-percentage increase in the current search index for the category "Real Estate Agencies" is associated with 16,550 additional sales for existing homes in contemporaneous quarter. Similarly, a one-percentage increase in the search index for the category "Real Estate Listing" is correlated with 23,360 houses sold in the present quarter (Model 3). While both contemporary search categories have explanatory power to predict the current home sales, predictions using the past search indices are mixed. The past search index on "Real Estate Listings is statistically correlated, albeit negatively correlated, with the present home sales (Model 2) while the past index on "Real Estate Agencies" is not. We explore the effect of using both the present and the past search indices for "Real Estate Listing" and "Real Estate Agencies" in Model 4. The present search indices for both categories are again positively correlated with sales, but the past

[7] We explored the following predefined categories on Google Trends: Apartments and Residential Rentals, Commercial & Investment Real Estate, Property Development, Property Inspection & Appraisals & Property Management, Real Estate Listings, Real Estate Agencies, Timeshares & Vacation Properties.

indices are not. However, the adjusted R^2 improved slightly when both the present and the past search indices are included. In Model 5, we only use the search indices to predict housing sales without lagged home sales and HPI and the results are similar to what is shown in Model 4. This suggests that using online search frequencies alone can predict future sales. The adjusted R^2 is just slightly below the baseline model when the past sales and the past HPI were included. Overall, results in Table 1 show that online search behaviors are highly correlated with the contemporaneous home sales.

Table 1. Linear Regression to Predict the Present Home Sales Using Search Frequency

Dependent Var.	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.864*** (0.0125)	0.864*** (0.0125)	0.819*** (0.0142)	0.842*** (0.0130)	0.806*** (0.0144))	
HPI_{t-1}	-0.140*** (0.0175)	-0.140*** (0.0175)	-0.158*** (0.0175)	-0.177*** (0.0196)	-0.188*** (0.0195)	
"Real Estate Agencies"$_t$		16.55*** (2.450)	17.09*** (3.424)		13.41*** (3.523)	48.47*** (6.415)
"Real Estate Agencies"$_{t-1}$			-0.780 (3.414)		1.170 (3.451)	33.04*** (6.297)
"Real Estate Listing"$_t$				23.36*** (4.797)	18.41*** (4.917)	37.37*** (9.007)
"Real Estate Listing"$_{t-1}$				-8.062* (4.831)	5.503 (4.876)	-13.16 (8.728)
Obs.	1561	1561	1561	1561	1561	1561
Controls	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population
States	51	51	51	51	51	51
Adjusted R^2	.973	.980	.981	.982	.983	.970

*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's

To examine whether our model can actually predict the contemporaneous home sales, we generate a set of a one-quarter-ahead predictions. We first create a training set using data from the 1st quarter of 2006 to the 4th quarter of 2008. Using these 11 quarters of data for 51 States, we select a set of features or variables that best predict the contemporaneous sales. We also experimented with various functional forms and the window of data to use that would give the best predictive results in the training set. We find a simple linear model with search terms to consistently provide superior prediction results. For predicting the present sales, using the previous 8 quarters of

data gives the best consistent results. In addition to using "Real Estate Agencies" and "Real Estate Listings", we also explored other pre-defined categories from Google Trends as well as our own bag of search phrases. However, we find "Real Estate Agencies" and "Real Estate Listings" are the best features for predicting the present sales in the training set. Next, we use the best-predicted model and estimates to predict sales from the 1st quarter of 2009 to the 3rd quarter of 2011. To gauge how accurate our predictions are compared to the actual real estate indicators, we use mean absolute error (MAE), which gauges how big the deviation the prediction is from the actual value. We can also use mean squared errors but the comparison does not qualitatively change from MAE.

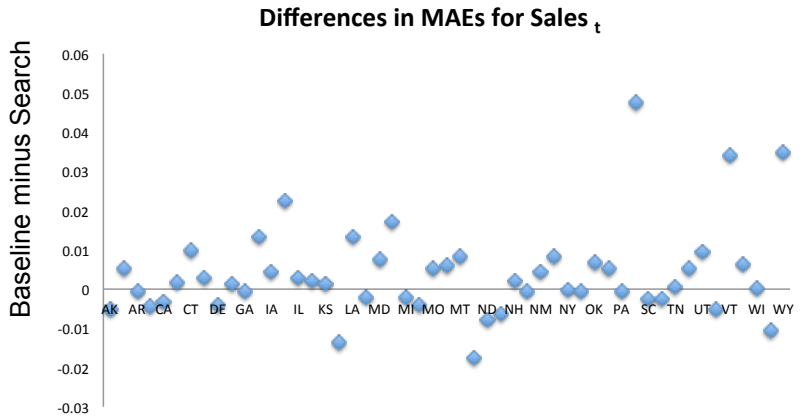


Figure 3: Y-axis indicates the average difference in MAE between the baseline model and the model that uses search indices. We use perditions from the first quarter of 2009 to the third quarter of 2011. When the dots are above the zero line, the baseline MAE is worse than the predicted MAE that uses search.

The mean absolute error (MAE) using Model 4 of Table 1 is 0.170, compared to 0.174, the MAE of the baseline model. Simply adding search terms in the linear model provides a 2.3% improvement over than the baseline (Model 0). We graphed the differences in MAE between the baseline model and the model that uses search indices in Figure 3, specifically as MAE(baseline) – MAE(search). When the dots are above the zero-line, predictions are better with search indices than the baseline model. However, when dots are below the zero-line, the baseline model performs better. As shown in Figure 3, the MAE for the baseline is mostly worse than our predictions that uses search. While the improvement is relatively modest on average, the variation for the improvement among different states is large. In general, predictions using search indices are better for states that have a high volume of sales,

possibly that high volume of sales also indicate that there are also more real estate related online searches. However, since search indices do not indicate the absolute number of searchers, so it is difficult to ascertain if more online searches leads to better predictions. We find the correlation between sales and the difference in MAE to be negative.

Table 2. Linear Regression to Predict the Future Home Sales Using Search Frequency						
Dependent Var.	Quarterly Sales $_{t+1}$	Quarterly Sales $_{t+1}$	Quarterly Sales $_{t+1}$	Quarterly Sales $_{t+1}$	Quarterly Sales $_{t+1}$	Quarterly Sales $_{t+1}$
	(0)	(1)	(2)	(3)	(4)	(5)
Sales $_{t-1}$	0.0864***	0.0696***	0.0687***	0.0807***	0.0656***	
	(0.00403)	(0.00452)	(0.00458)	(0.00419)	(0.00465)	
HPI $_{t-1}$	0.959***	0.952***	0.951***	0.952***	0.947***	
	(0.00558)	(0.00556)	(0.00557)	(0.00626)	(0.00620)	
"Real Estate Agencies" $_{t}$		5.986***	5.069***		3.520***	6.817
		(0.780)	(1.107)		(1.138)	(4.543)
"Real Estate Agencies" $_{t-1}$			1.268		2.361**	9.146**
			(1.088)		(1.104)	(4.414)
"Real Estate Listing" $_{t}$				8.951***	7.919***	16.82***
				(1.528)	(1.560)	(6.246)
"Real Estate Listing" $_{t-1}$				-5.116***	-4.989***	51.97***
				(1.514)	(1.523)	(5.945)
Obs.	1561	1561	1561	1561	1561	1561
Controls	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population
States	51	51	51	51	51	51
Adjusted R^2	0.971	.976	.978	.978	.979	.965
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

Next, we apply our methods to predict the future housing trends using available data today that include the past housing statistics and the present search indices. We can only use the past housing statistics from the previous quarter because the present housing statistics has not been released yet. Unlike housing statistics, which is always released with a lag, search indices are obtainable instantaneously, allowing us to incorporate real-time search behaviors to predict future real estate trends. We show the correlations between search indices and future housing sales in Table 2, which largely supports our hypothesis that search indices can be used to predict future housing sales in the future. The search category "Real Estate Agencies" is positive and statistically significantly correlated with future home sales. As shown in Model 4, a one-percentage increase in the search index for "Real Estate Agencies" in the previous quarter is associated with additional 3,520 units of future sales in the future quarter.

Similarly, the present search index for "Real Estate Listing" is also positively correlated with future sales (β=7.919, $p<0.01$). Overall, adding the search terms improved the fit of the model, providing evidence that search indices can be used to predict future home sales.

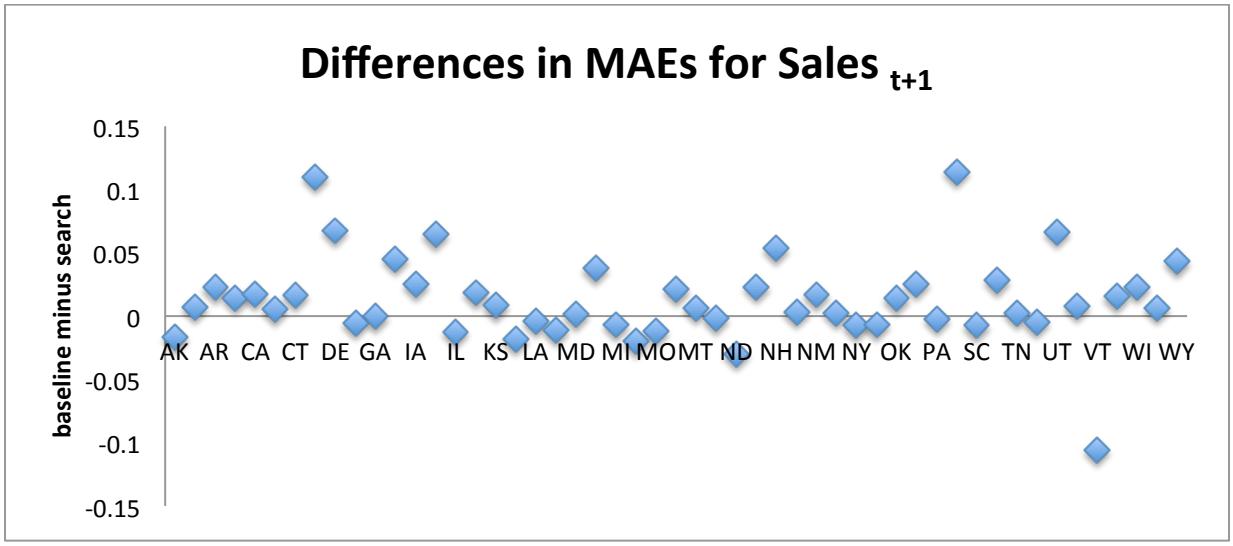


Figure 4: MAE differences between the baseline model and predictions using search indices.

Next, we explore if search indices can actually predict future home sales. From the training data, we find the best functional form and variables include in the model. We also experimented with how much historical data to include for predictions and find that using a window of 8 quarters immediately before the present to best predict the present and future sales. After experimenting with various housing related search terms and predefined search categories from Google Trends, we find the best predictors are the current index for "Real Estate Agencies" as well as its one-quarter and two-quarter lags. Interestingly, the "Real Estate Listings" index no longer adds much predictive power when indices on "Real Estate Agencies" are included. Using only the present and the past indices on "Real Estate Agencies" as well as the past statistics on HPI and home sales, we predict the future home sales and plot the difference between the MAE of the baseline model and the MAE of our predictions in Figure 4: MAE(baseline) – MAE(search). For most of the states, predictions using search indices outperform the baseline predictions, especially for states where the sales volume is high. For states with lower real estate transactions, adding search indices does

not improve the baseline predictions. Overall, the MAE for predictions using search indices is 0.172 while the baseline MAE is 0.185. This is a 7.1% improvement over the baseline model. Interestingly, this result suggests that search indices are better at predicting the future home sales than the contemporaneous sales (7.1% vs. 2.3% over the baseline). Perhaps, future sales are more correlated with past search indices because buying and selling a house often take a long time. For example, while there are many factors affecting the duration of a sale, the average time to sell a home in the US is 10 months in 2011[8]. Thus, search activities on the Internet can potentially forecast home sales 10 months in the future. This type of behavioral information could be more valuable for predicting the future than information provided by the two-period lags of home sales and HPI.

While we find that using search indices can improve prediction outcomes from the baseline model, it is important to compare our model with real forecasts from experts in the field. Thus, we have collected data from the National Association of Realtors that release quarterly forecasts for US home sales. We compared their forecasts with our predictions from the second quarter of 2009 to the third quarter of 2011, for a total of 10 quarters. When predicting the present home sales, we find that our predictions are slightly better than NAR's but the difference is not statistically significant. However, our predictions are considerably better than NAR for predicting future home sales. The MAE for the National Association of Realtor's forecast is 0.110 while the MAE for model that uses search indices is 0.084, a 23.6% improvement over the estimates from real estate experts. Results are summarized in the Table 3 below. This again shows the power of using search indices for predicting the future. Using a simple linear prediction model with search indices, we are able to outperform predictions from established experts in the field.

Table 3: Comparing with predictions with the National Association of Realtors

MAE for Sales $_{t+1}$	Obs.	Mean	Std Err	Min	Max
Search	10	**0.084**	0.0316	0.0122	0.1556
NAR	10	**0.110**	0.0262	0.0504	0.1688
Diff		**23.6%**	P<0.01		

[8] Statistics from the Accredited Seller Agent Council. http://www.realty101.com/what-is-the-average-time-to-sell-a-home

Predicting the House Price Index Using Online Search Data

In Table 4, we explore the relationship between the housing related search indices and the house price index (HPI), which is calculated based on a modified version of the weighted-repeat sales (WRS) methodology proposed by Case and Shiller (1989). Similar to models in Table 1, all the models in Table 4 use a fixed-effect specification on an AR model with region, population and seasonality controls. As expected from the baseline AR model (Model 0), the lagged HPI and lagged sales are positively correlated with the present HPI. In Model 1, we estimate the correlation between the current search index for "Real Estate Agencies" and the HPI and find that a one-percentage increase in search index is associated with an increase of 5.986 points in HPI. However, the past search index on "Real Estate Agencies" from the previous quarter does not have statistically significant correlation to the present HPI (Model 2). Next, we introduce both the current and the past indices of for "Real Estate Listings" in Model 3. We find that the current search index for "Real Estate Listings" is positively correlated with the contemporaneous HPI while its one period lag is negatively correlated with HPI. Finally we include the present and the past search indices for both "Real Estate Listings" and "Real Estate Agencies" in Model 4 and find that all search indices are correlated with the present HPI. The fit of model also improves slightly. These results give us confidence that incorporating the present and the past search indices from the two search categories can help predicting the contemporaneous HPI.

Next we predict the contemporaneous HPI from the 1st quarter of 2009 to the 3rd quarter of 2011 after finding the best-fitted model from the training data set. In addition to "Real Estate Agencies" and "Real Estate Listings", we also explored other real estate categories as well as their lags. We continue to find the contemporaneous and the one-period lag of search indices on "Real Estate Agencies" and "Real Estate Listings" to best predict the present HPI. In contrast to using the previous 8 quarters of data to predict home sales, we find that using data from the past four quarters can best predict the present HPI. Overall, we find that our predictive accuracy improve from the baseline model by 2.54%, which is comparable to the results on predicting the present home sales. We show the state-by-state scatter plot for the MAE difference between the baseline and the search indices model (Figure 5). Again, dots above the zero-line represents states where the prediction using search is superior than the baseline while the opposite is true for dots below the zero-line.

Table 4: Linear Regression of HPI on the Search Index Related to Real Estate and Real Estate Agencies						
Dependent Var.	HPI_t	HPI_t	HPI_t	HPI_t	HPI_t	HPI_t
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.959***	0.952***	0.951***	0.952***	0.947***	
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	
HPI_{t-1}	0.0864***	0.0696***	0.0687***	0.0807***	0.0656***	
	(0.004)	(0.0052)	(0.005)	(0.004)	(0.005)	
"Real Estate Agencies"$_t$		5.986***	5.069***		3.520***	6.817
		(0.780)	(1.107)		(1.138)	(4.543)
"Real Estate Agencies"$_{t-1}$			1.268		2.361**	9.146**
			(1.088)		(1.104)	(4.414)
"Real Estate Listing"$_t$				8.951***	7.919***	16.82***
				(1.528)	(1.560)	(6.246)
"Real Estate Listing"$_{t-1}$				-5.116***	-4.989***	51.97***
				(1.514)	(1.523)	(5.945)
Obs.	1561	1561	1561	1561	1561	1561
Controls	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population
States	51	51	51	51	51	51
Adjusted R^2	0.987	0.986	0.987	0.987	0.987	0.987
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

Overall, using search, we are able to predict 39 states better than the baseline model, but our predictions are particularly worse for a few states such as Alaska, Montana and South Dakota. These states tend to have fewer transactions on housing sales than other states. Similar to what we found for home sales, search indices help predictions the most when sales volume is high. Furthermore, predicting HPI may just be inherently more difficult than predicting home sales. While home sales can increase when either the housing demand or supply changes, HPI would increase only when the demand for housing is increased but decrease when the supply is increased. It can be very difficult to know whether the search queries in general categories such as "Real Estate Agencies" or "Real Estate Listings" are coming from the demand side or the supply side, and thus it can be much harder to predict HPI than the volume of home sales. For example, both sellers and buyers need real estate agents, so an increase in the search index related to real estate agencies could come from both the supply and the demand sides. Thus, it is difficult to make a linear prediction between the search index and HPI. To address this issue, we tentatively aggregated some search terms relating to buyers activities such as home financing, mortgage and home inspections and also some search terms related to seller's activities only such as home staging. Currently the indices we created

ourselves are not powerful enough to accurately predict future HPI. However, given how fine-grained the search terms are, we are hopeful to create search indices that can distinguish movements from the demand side away from the supply side.

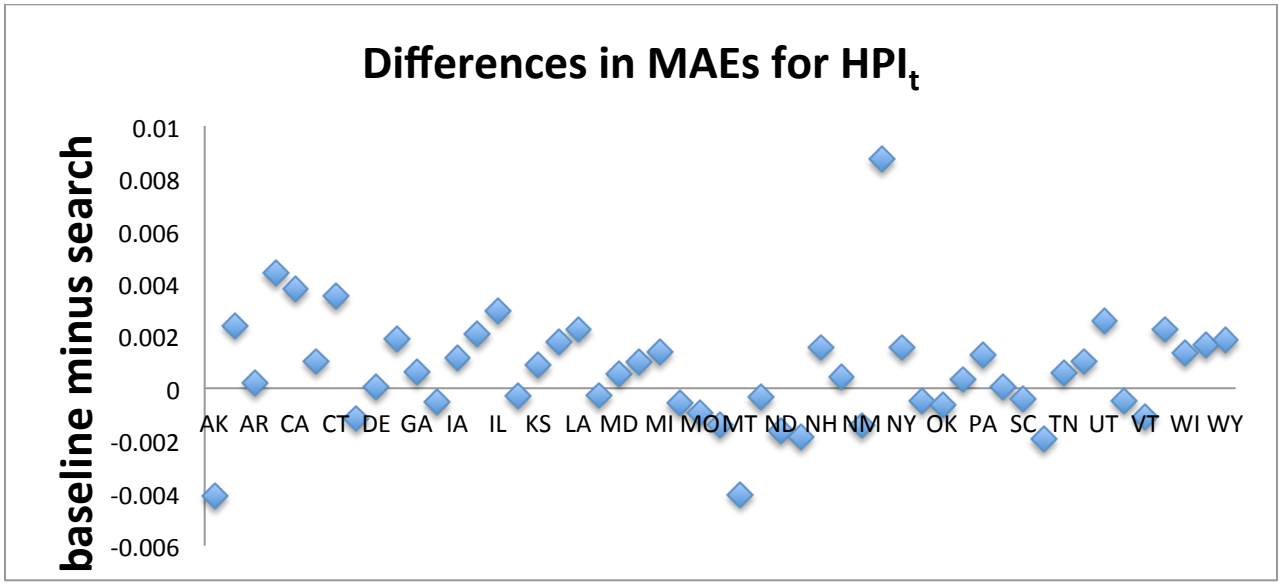


Figure 5: Difference in MAE between the baseline model and the search-based model

To explore how search indices can be used to predict future HPI in the next quarter, we first show whether these search terms are correlated with the future HPI. In Model 0 of Table 4, we estimate the baseline AR model and show that the future HPI is correlated with the both the past HPI as well as the past sales from the previous quarter. In Model 1, we incorporated the current search index for "Real Estate Agencies", and find that the current search index is correlated with the future HPI in the next quarter (β=5.986, $p<0.001$). In Model 2, we incorporate both the current and the past search indices to predict the future HPI. We find while the current index on "Real Estate Agencies" is correlated with the future HPI, the search index in the previous quarter is not. We the explore how search indices on "Real Estate Listings" are correlated with future sales and find both the past and the present search indices are correlated with forecast future sales in the next quarter (Model 3). Lastly, we used search indices for both "Real Estate Agencies" and "Real Estate Listings" in Model 4. While all the search indices were statistically significant in predicting the future HPI in the next quarter, the over fitness of the model is roughly constant. When we use only the

search indices to predict the future HPI (Model 5), the adjusted R^2 significantly dropped compared to the previous models including the baseline model, even though all search indices continue to be statistically significantly correlated with the future HPI. These results highlight the difficulty of forecasting HPI, as the search volume can potentially shift both the supply and the demand curve simultaneously.

Table 4: Linear Regression of HPI on the Search Index Related to Real Estate and Real Estate Agencies

Dependent Var.	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.0864*** (0.004)	0.0696*** (0.005)	0.0687*** (0.005)	0.0807*** (0.004)	0.0656*** (0.005)	
HPI_{t-1}	0.959*** (0.006)	0.952*** (0.006)	0.951*** (0.006)	0.952*** (0.006)	0.947*** (0.006)	
"Real Estate Agencies"$_t$		5.986*** (0.780)	5.069*** (1.107)		3.520*** (1.138)	6.817 (4.543)
"Real Estate Agencies"$_{t-1}$			1.268 (1.088)		2.361** (1.104)	9.146** (4.414)
"Real Estate Listing"$_t$				8.951*** (1.528)	7.919*** (1.560)	16.82*** (6.246)
"Real Estate Listing"$_{t-1}$				-5.116*** (1.514)	-4.989*** (1.523)	51.97*** (5.945)
Obs.	1561	1561	1561	1561	1561	1561
Controls	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population	Quarters, States, Regions, Population
States	51	51	51	51	51	51
Adjusted R^2	0.986	0.987	0.987	0.987	0.987	0.786

*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's

Next we use the training data to find the best features that can be used to predict the future HPI. In addition to using the present and past search indices of "Real Estate Agencies" and "Real Estate Listings", we also explored some nonlinear forms of the search indices such as their quadratic terms. We also created our own search indices based on search keywords that could shift the supply curve and also those that could shift the demand curve. Perhaps, our current keyword selection is still too crude to tease out the differences between the supply and the demand. Overall, we find the best predictors continue to be the present and past indices for "Real Estate Agencies" and "Real Estate Listings." In addition, we find the quadratic terms of "Real Estate Agencies" to also help with the predictive accuracy in the training set. Thus, we include these variables to predict the future HPI from the 2nd quarter of 2009

to the 3rd quarter of 2011. We plot the difference in MAE between the baseline model and the search model for each state of the US in Figure 6. For most states, predictions using search were better than the baseline model, though the variance among states is higher than previous predictions. We predicted 11 quarters for 51 states. The baseline MAE is 0.027 and the MAE using search is 0.0262, about a 2.96% improvement in accuracy and the difference is statistically significant at p = 0.01 level. Unfortunately, the National Association of Realtor does not forecast HPI, at least from public available sources and thus, we are not able to compare our predictions with NAR's.

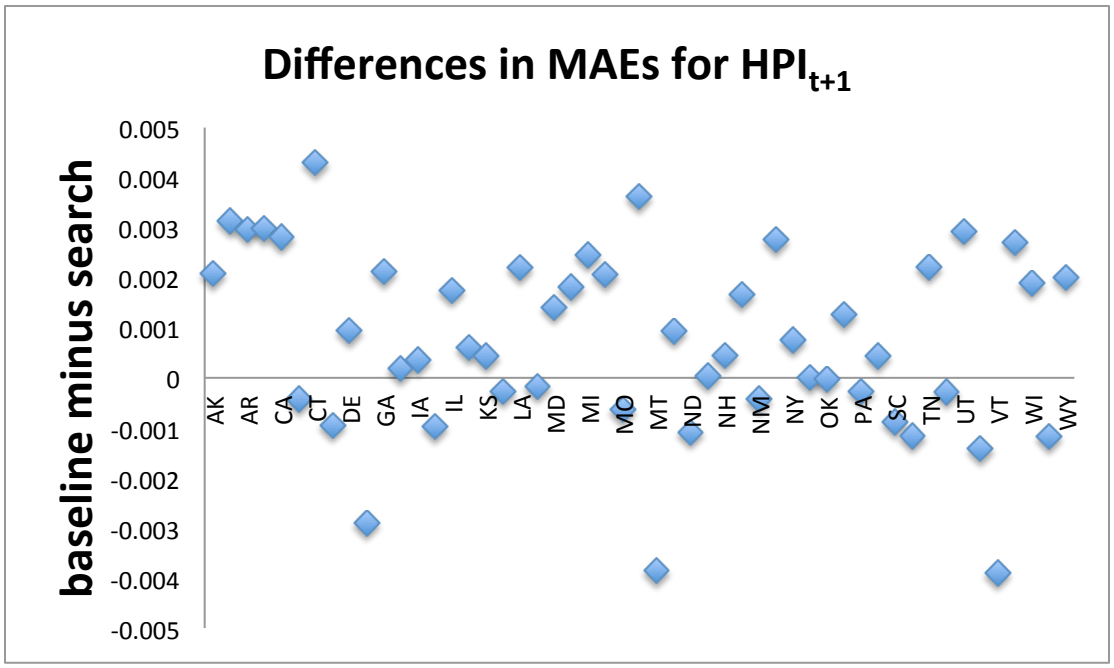


Figure 6: The MAE difference between the baseline model and our prediction model.

We summarize our results in Table 5. While using search frequencies can more accurately predict the present and future home sales as well as HPI, it is more effective for predicting the future than predicting the present. Perhaps, because the nature of the housing transaction that often takes months to more than a year to complete, search indices in the present can be particularly useful to forecast future housing indicators. We also find that it is relatively easier to predict the sales volume than the HPI. As we have explained earlier, predicting the price can be inherently more

difficult than predicting the volume because predicting price requires the ability to distinguish housing supply from housing demand. While an increase in housing demand can increase home price, a similar increase in supply would actually decrease the price. We also explore how using search frequencies can distinguish the supply from the demand in the housing market by refining the search categories to differentiate searches that are buyer specific from those that are seller specific. For example, home buyers are more likely to look for loans than sellers whereas sellers are more likely to hire a staging company to make the property more appealing to the highest number of potential buyers. We would therefore expect that an increase in search frequencies related to financing and loans to shift the demand curve while a similar increase for searches related to home staging is more likely to shift the supply curve for housing. We see some evidence that home financing are positively correlated with HPI, suggesting it may be shifting the demand outward. Currently, we have not found a set of queries that could shift the supply curve. However, because of the fine-grained nature of the search term, we are hopeful that indices can be created to precisely tease out a shift in the demand curve from a shift in the supply curve.

Table 5. Summary of MAE for predicting the present and the future housing trends

	Obs.	MAE_search	MAE_baseline	% Improvement over baseline
$Sales_t$	561	.170	.174	2.3%**
HPI_t	561	.026	.027	2.45% ***
$Sales_{t+1}$	561	.172	.185	7.1% **
HPI_{t+1}	561	.026	.027	2.96%***

Predicting the Demand for Home Appliances

Lastly, we explore trends in home appliance sales. We expect that housing sales would spur interests in buying home appliances, increasing their demand in the future. To gauge the overall interests in home appliances, we use the search index for the "Home Appliance" category from Google Trends and show its relationship with home sales (Table 6). We observe that the current home sales are not correlated with the contemporaneous search index for home appliances (Model 1, Model 4). But after 6 months, each one thousand houses sold previously is correlated with a 1.14 percentage point increase in the search index for home appliances. Since buyers move into their new

properties first before making major purchases (and often researching such purchases), it is natural that the number of online searchers for home appliances would only increase after a consumer has already bought a house. Thus, we may expect the online search for home appliances to lag behind housing sales. The actual demand for home appliance may rise after this increase in the appliance search index if some of the online searches translate into future sales. Similarly, we correlated the housing real estate related search index with the home appliance search index and we find that they are also positively correlated (Column 2 Table 6) This highlights the linkages between home sales and other parts of the economy that may complement real estate purchases.

Table 6. Linear Regression on Search Terms Related to Home Appliances and the Volume of Housing Sales

Dependent Var. Search terms related to home appliances	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)
	(1)	(2)	(3)	(4)
	Fixed effect	Fixed effect	Fixed effect	Fixed effect
Home Sale Vol	-.054 (.00011)			0.188 (0.000359)
Home Sale Vol – lagging 1 quarter		-.02 (.00014)		-0.627 (0.393)
Home Sale Vol – lagging 2 quarters			.59** (.3)	1.14*** (0.427)
Obs.	254	203	152	152
Controls	Quarters	Quarter	Quarters	Quarters
States	51	51	51	51
*p<.1, **p<.05, ***p<.001, Huber-White robust standard errors are shown in parentheses				

Implications

Twenty-five years ago, Herbert Simon (1984) observed:

> "In the physical sciences, when errors of measurement and other noise are found to be of the same order of magnitude as the phenomena under study, the response is not to try to squeeze more information out of the data by statistical means; it is instead to find techniques for observing the phenomena at a higher level of resolution. The corresponding strategy for economics is obvious: to secure new kinds of data at the micro level"

Today, advances in information technology in general, and Internet search query data in particular, are making Simon's vision a reality. Who could have imagined that we would be observing literally billions of consumer and business intentions to buy or sell before they even occur in the marketplace? Yet, that is what search query data

does. What's more, we can do so at nearly zero cost, virtually instantaneously and at remarkably fine-grained levels of disaggregation. These data are increasingly available to ordinary consumers, business people and researchers of all types.

We have found that analyzing online search data with relatively simple models can yield more accurate predictions about the housing market than were previously possible. If online search patterns can be construed as a broad indicator of interest within a group, it can also be used as a reliable predictor to forecast economic activity. Analyzing housing market trends, we find evidence that search indices adds substantial power in predicting the underling economic trends and predictions using search indices can outperform predictions from experts in the field such as the National Association of Realtors. They support to the hypothesis that Web search can be used to predict present and future economic activities. For example, housing-related search can be used to predict the turning point in the economic recovery from a recession.

Currently, we are able to make fairly accurate predictions using simple linear prediction model using a few predefined real estate categories in Google Trends. Because of the fine-grained nature of these data, they can be aggregated in many different ways to predict specific underlying economic shifts. For example, instead of using rough categories such as "Real Estate Agencies" or "Real Estate Listings", we can create our own bag of words specific for gauging changes in demand as well as changes in supply. For example, search phrases such as "mortgage", "finances", "school district" are more likely to come from buyers while phrases such as "home staging" and "professional cleaner", "contractors" are more like to come from sellers. Distinguishing the search indices from the supply side from the demand side can more accurately gauge what is driving the change in real estate market. Similarly, we can test more fine-grained predictions about the real estate market beyond sales and price. For example, search indices can be create to gauge the interest of people buying real estate as opposed to renting or whether new construction activities are growing over time or not. Because the fine-grain nature of individuals' search queries, it is possible to construct different types of indices and quickly test their validity in predicting various real estate trends and beyond.

Timely and accurate predictions about the housing market can benefit a wide array of industries, such as construction and home appliances, as well as individuals, such as homebuyers and sellers. Since buying a home is

the single biggest expenditure and one of the biggest financial decisions for most people, obtaining accurate and timely information can help them make informed decisions and potentially save tens of thousands of dollars for the average family. Similarly, businesses that depend on the housing market can benefit from this simple use of Internet search data. Currently, economists and investors primarily rely on housing data released from the government and trade groups such as the National Association of Realtors, to understand the current housing market and forecast future market trends. However, government and trade group data are released with a delay and often with pending revisions. Furthermore, they do not provide fine-grained reports at the town level that is crucial for buyers and sellers to make informed decisions. With easy access to billions of online search frequencies, it is now possible to use a simple technology to cheaply collect timely, accurate and fine-grained analysis about the housing market. Not only does Google Trends provide weekly reports on the volume of housing related queries, it also offers a detailed regional analysis at country, state and city levels. By leveraging micro data collected from Google Trends, investors can obtain deeper insight about the housing market in order to make informed decisions.

Accurate predictions on the housing market can also have strong ripple effects on other sectors of the economy, especially for its complementary goods. For example, timely and accurate forecasts of housing demand allow the construction industry to improve future plans for developments and thus reduce the probability of experiencing the housing boom and bust cycles. Similarly, accurate housing market forecasts can also help the home appliances industry to manage its inventory.

Other applications of this research

While we show promising predictions about the housing market using Google Trends, it can be also used in many other contexts to predict future economic activities, for example, the technology sectors. In particular, Google trends can be used to predict the outcome of the standards war. We were able to track the progression of the standards war between HD-DVD and Blue-ray to play out on Google Trend and search indices were quite prescient in predicting that Blue-ray would win in the end. Similarly, we can also use search frequency to predict the market share of an electronic product or an operating system such as Macintosh. Instead of paying a premium for industry reports, Google Trends can be used to predict if a particular technology would gain market shares.

Conclusions, Limitations and Future Work

Four hundred years ago, the microscope was invented. For the first time, scientists could see individual microbes in a drop of water and blood cells that traversed through the body. The result was a revolution in biology and medicine, including the germ theory of disease and ultimately, new vaccines, medicines and treatments. Today, due to advances in IT and IT research, we are gaining the capability to observe micro-behaviors online. Rather than rely on painstaking surveys and census data, predefined metrics and backward-looking financial reports, social science researchers can use query data to learn the intentions of buyers, sellers, employers, gamers, gardeners, lovers, travelers and all manner of other decision-makers even before they execute their decisions. It is possible to accurately predict what will happen in the market place days, weeks and even months in the future with this approach. Search technology has revolutionized many markets, and it is now revolutionizing our research.

This is an exploratory study investigating whether online search behavior from Google Search can predict underlying economic activity. Using housing sales data, we find evidence that search terms are correlated with sales volume and also with the house price index, lending credibility to the hypotheses that Web search can be used to predict future economic activity, for example when the economy may recover from the current recession. We are aware of the fact Google search queries do not represents all the online housing search activities. As some consumers may bypass the search engine all together and go directly to certain websites such as Realtor.org when considering buying and selling a home. Approach using Google search alone would miss this type of consumers. However, despite missing these consumers, we can still predict the housing sales and housing price using only online search captured by Google, demonstrating the power of online queries in forecasting economic trends.

Ultimately, micro data collected using Google Trends may prove one of the most powerful tools for helping consumers, businesses and government officials make accurate predictions about the future so that they can make effective and efficient decisions. It distills the collective intelligence and unfiltered intentions of millions of people and businesses at a point in their decision-making process that precedes actual transactions. Because search is generally not strategic, it provides honest signals of decision-makers intentions. The breadth of coverage, the level of disaggregation and the speed of its availability is a radical break from the majority of earlier social science data. Even simple models can thus be used to make predictions that matter.

Of course, there are many obstacles yet to overcome and refinements to be made. For instance, paradoxically, as businesses and consumers come to rely on query data for their decision-making, as we expect they will, there will be incentives for opposing parties to try to degrade the value of the data, perhaps by generating billions of false or misleading queries. This will in turn call for counter-measures and perhaps the golden age of simple models using these data will be brief. However, more than four years have passed since we first started using Google Trends to forecast real estate trends. Surprisingly, search indices continue to have the power to predict the future, as we have shown in this paper. Informational value derived from search indices has not been absorbed into economic equilibria as many have argued. Instead, its effect, at least for the real estate market, has persisted over time. Meanwhile, new types of nanodata have become available, such as Twitter feeds, social networking data and various forms of digital trace. Along with search, detailed nano-data has continued to proliferate at a pace that has far outgrown our ability to manage and use these data appropriately. For example, a simple hoax using a single Twitter feed in April 2013 has erased Dow Jone's industrial by 145 points in less than five minutes. Perhaps because of the short nature of the stock price fluctuation, a fake Twitter about a bombing in the White House can quickly go viral and affect the trading strategies of many high frequency traders. Consequently, the stock market erased $136 billion in equity in mere three minutes. However, this type of gaming is less likely to happen for market that is not prone to change so quickly, such as buying and selling a house. Because the nature of selling a home can take months to complete, a swing in search indices on housing queries in an hour or a day would not make a significant impact on the predictions of future real estate trends. These types of hacking are often quickly discovered using tests for statistical anomalies. Thus, making long-term manipulation becomes more difficult. Future research should investigate on what types of market search and other forms of digital trace are most useful for predictions and what types of markets are susceptible to gaming. We have so far identified that the speed of the market change may play a role but many other factors could also be at play.

Ultimately, the availability of various digital trace has grown so quickly over years that they have outpace our ability to understand and use them effectively. It is thus important for future research to investigate how to integrate and use them in a meaningful fashion to understand underlying consumer sentiments and economic consequences. Through better understanding, we may be able better distinguish malicious and faulty information from the true economic signals, although it may be also be a cat-and-mouse game where malicious attack will always happen on strategic tools that can affect our decision-making. Through these explorations, we will also have a better

understanding of what types of market can benefit from the use of nanodata in predictions and what types of markets are more difficult. Perhaps some markets require higher data quality and some are just prone to be manipulated, such as the stock market. There might be some predictions that will always be difficult to do regardless of how fine-grained data have become. For example, predicting when earthquake will strike at any particular moment in time may just be inherently more difficult and data on human behaviors may play minimal roles in the prediction. However, as more nanodata and methods become more widely used, we can only conclude that the future of prediction is far brighter than it was only a few years ago.

References

Appleton-Young, L., 2008, "State of the California Housing Market 2008-2009, California Association of Realtors

Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2006. "Information, Technology and Information Worker Productivity: Task Level Evidence." Proceedings of the 27th Annual International Conference on Information Systems, Milwaukee, Wisconsin.

Calhoun, C.A, (1996) "OFHEO House Price Indexes: HPI Technical Description", OFHEO, http://www.fhfa.gov/webfiles/896/hpi_tech.pdf

Case, K.E. and Shiller, R.J. (1987). "Prices of Single Family Real Estate Prices," New England Economic Review. 45-56.

Case, K.E. and Shiller, R.J. (1989). "The Efficiency of the Market for Single-Family Homes," The American Economic Review. 79, 125-137.

Davenport, T. (2006) "Competing on Analytics" Harvard Business Review Article, Jan, 2006

Choi, H., Varian, H., 2009 "Predicting the Present with Google Trends, April 2009", http://www.google.com/insights/search/#

Horrigan, "The Internet and Consumer Choice", Pew Internet and American Life Project, May 2008. http://www.pewInternet.org/~/media/Files/Reports/2008/PIP_Consumer.Decisions.pdf.pdf

Glaeser & Gyourko, 2007, "Housing Dynamics", NBER Working Paper

Glaser, 2009 "Housing Prices in the Three Americas", http://economix.blogs.nytimes.com/2008/09/30/housing-prices-in-the-three-americas/

Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant, "Detecting influenza epidemics using search engine query data", Nature vol. 457, November 2008.

Han, L (2008) Hedging House Price Risk in the Presence of Lumpy Transaction Cost, *Journal of Urban Economics (64)* , February 2008, 270-287

Han, L., (2009) "The Effects of Price Uncertainty on Housing Demand: Empirical Evidence from the U.S. Markets", Working Paper

Krugman, P., "How Did Economists Get It So Wrong?", New York Times, September 2, 2009, http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html?em

Kuruzovich, J., Viswanathan, S., Agarwal, R., Gosain, S. and Weitzman, S. 2008. "Marketspace or Marketplace? Online Information Search and Channel Outcomes in Auto Retailing," Information Systems Research 19:2, pp. 182-201.

Moe, W. W., and Fader, P. S. 2004. "Dynamic Conversion Behavior at E-Commerce Sites," Management Science 50:3, pp. 326-335.Pentland, A., "Honest Signals: How They Shape Our World", The MIT Press: London, 2008

National Association of Realtors, "Profile of Home Buyers and Sellers", 2012

comScore, 2012, http://www.comscore.com/Insights/Press_Releases/2013/1/comScore_Releases_December_2012_U.S._Search_Engine_Rankings, December, 2012

New York Times Editorial, "Unemployment Rising", http://www.nytimes.com/2009/04/05/opinion/05sun1.html, April 4, 2009,

Pentland, A., "Honest Signals", MIT Press, 2008

Simon, Herbert A. "On the Behavioral and Rational Foundations of Economic Dynamics." Journal of Economic Behavior and Organizations, Vol. 5, (1984), pp. 35-66.

Wu, L., Waber, B., Aral, S., Brynjolfsson, E., & Pentland, A. ""Mining Face-to-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an IT Configuration Task", International Conference on Information Systems, Paris, France, December 14 – 17, 2008.