2009

The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities

Lynn Wu
Massachusetts Institute of Technology, lynnwu@mit.edu

Erik Brynjolfsson
Massachusetts Institute of Technology, erikb@mit.edu

THE FUTURE OF PREDICTION: HOW GOOGLE SEARCHES FORESHADOW HOUSING PRICES AND QUANTITIES

Completed Research Paper

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

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

Abstract

To make effective decisions, consumers, executives and policymakers must make predictions. However, most data sources, whether from the government or businesses, are available only after a substantial lag, at a high level of aggregation, and for a small set of variables that were defined in advance. This hampers real-time prediction. A critical advance in IT research has been the development of powerful search engines and the underlying Internet infrastructure. We demonstrate a highly accurate but simple way to predict future business activities by using data from such search engines. Applying our methodology to predict housing trends, we find that our index of housing search terms can predict future quantities and prices in the housing market. During our sample period, each percentage rise in our housing search index predicts sales of 121,400 additional houses in the next quarter. This approach can be applied to other markets, transforming the future of prediction.

Keywords: Online Search, Housing Index, Housing Trends, Predictions

"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 with significant delay and aggregated into a relatively small number of pre-specified categories. This severely 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 trillions 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 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 Great 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 fundamental transformation on how we solve the intractable problem of predicting the future. Just as the capacity to collect and analyze massive data has transformed fields in life sciences, IT acts as the catalyst to revolutionize social science research by making available massive amount of data at an extraordinarily fine-grained level, so that researchers could tackle hard questions that were impossible to answer before. With the observation of trillions of consumers and business intentions as revealed by online search, we show that we can yield surprisingly accurate predictions about future economic activities using a few simple techniques.

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 market place. 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, the vast majority of the economy. What's more, by identifying correlations with prices and quantities we can make inferences about changes in 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 activities. Using aggregation of 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[1].

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.

[1] Our preliminary work shows that Google Trend can be used to forecast the winner of competing standard such BlueRay vs. HD-DVD as well as mainstream operating systems.

In turn, when the housing market becomes healthy again, the recession may also come to an end (New Work Times, 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 end of the current recession. 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 predict the future housing price via the frequency of online search terms submitted to a search engine. Analyzing consumers' interests as revealed by their online behaviors, we are able to uncover sales trends before they even appear.

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 2008 Profile of Home Buyers and Sellers by National Association of Realtors (NAR), 87% of homebuyers used the Internet to search for a home in 2008 (NAR, 2008). Similarly, a report, written by California Association of Realtors in 2008, shows that 63% of home buyers find their real estate agent using a search engine (Appleton-Young, 2008). Clearly, the Web has become an integral element for searching to buy a home. If we can discover a mechanism that can capture online search activities, we may be able to link the current housing inquires to future housing sales. To test this hypothesis, we rely on hundreds of billions of individual searches from five years of the Google Web Search portal[2] to predict housing sales and housing prices. Using these detailed 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 the search term frequency can be used to predict future housing sales. Specifically, we find that a 1% increase in search frequency about real estate agent is associated with quarterly sales of an additional 67,700 homes in a US state. This is substantial as the typical State in the US sells on average 96,339 units of houses per quarter[3].

Similarly, we also examine the relationship between housing price and housing related online search. Using the house price index (HPI) from Federal Housing Finance Agency[4], we find a positive relationship between the housing related online queries and the present HPI. This appears to reflect an increase in housing demand, driven by home buyers who search for houses online prior to actually buying. Interestingly, the HPI is negatively correlated with housing queries three months prior. We infer this to correspond to an increase in the supply of available houses in the market. Sellers "move first" in this marketplace, surveying the competition and assessing market conditions before making a decision to sell. As more sellers reveal their intentions, more houses eventually become available for sale. In turn, the listing price is likely to fall, driving down the overall house price index.

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 on queries about home appliances as a measure for consumers' interests for the product (Moe and Fader, 2004) and find that every thousand houses sold is correlated with a 1.23% 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 may complement home sales.

Literature Review

Predicting future social and economic trends have always been an important contribution to social science. In the past decades, much of the social science research has focused on building complex mathematical models to predict the future. However, collecting high quality data at a more fine-grained level has mostly eluded social science research. Yet the need to obtain such data has become growingly important for businesses as well as social science research.

---

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

[3] Calculated using data from the National Association of Realtors

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

Today, advances in information technologies, such as the Internet search technologies, e-mail, smart sociometric badges, 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 a micro-level. Similarly, Aral et al. (2007) used email data to capture real-time communication patterns of a group of recruiters over several years. They were able to examine work behaviors, such as multitasking, and their impact on long term work performance. Lazer et al (2009) has provided various examples of using high quality data, produced by new technologies, that has the potential to transform social network research. Similarly, firms have also leveraged large-scale data collected online to make predictions, such as consumer preferences, supply and demands for various goods as well as basic operational parameters, such as inventory and turnover rate. The ability to collect and efficiently analyze the massive amount of data made available by information technology has enabled firms, such as Amazon, Harrah's and Capital One, to achieve tremendous gain in profitability and market shares (Davenport, 2006).

Our work follows a similar stream in demonstrating the power of using accurate and fine-grained data to predict underlying social and economic trends. Unlike previous research and business analyses that have primarily used proprietary data, we leverage free and public available data from Google to accurately forecast economic trends. Significant work has shown that online behaviors can be used to reveal consumers' intention and can be used to predict purchase outcomes (e.g. Moe and Fader, 2004; Kuruzovich et al. 2008). Based on the ability of online activities to reveal human behaviors, we 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.

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 future economic trends, such as forecasting price and quantity of houses sold in the future. 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 by geographical locations. Furthermore, we use search frequencies to infer the separate shifts in the underlying supply and demand for the real estate market.

Data Source

Google Search Data

We collected the volume of Internet search queries about real estate from Google Trend which provides weekly and monthly reports on query statistics related to various industries. Furthermore, Google Trend allows users to obtain the volume of queries for a specific phrase, such as "Housing Price". Since 2004, Google Trend has systematically captured online queries submitted to the Google Search Engine 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 most widely used search engine, which processed more than 60% of all the online queries in the world (Nielsen Report, 2008), it is possible to approximate people's interests over time using the volume of queries submitted to Google. In fact, recent work using Google Trend can accurately predicts flu outbreaks few days before it actually happens (Ginsberg et al, 2008). We believe that Google's search volumes can also be used to predict future economic indicators.

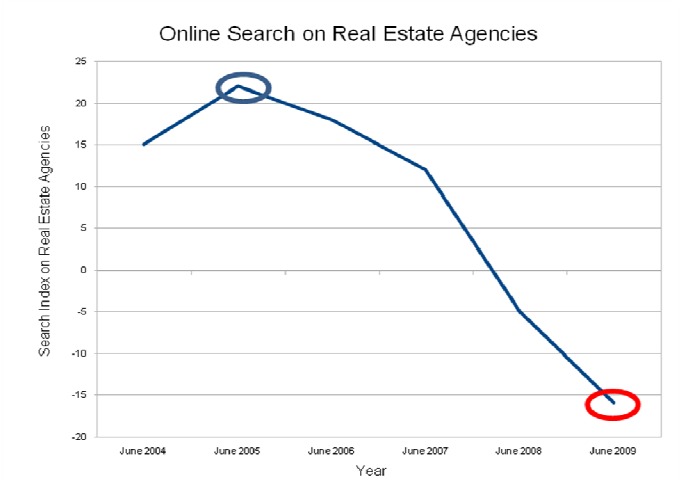

Figure 1: Search Index for "Real Estate Agencies. It is a normalized measure of search volume on the categories of search phrases related to real estate agencies. The index is always a percentage from 0 to 100.

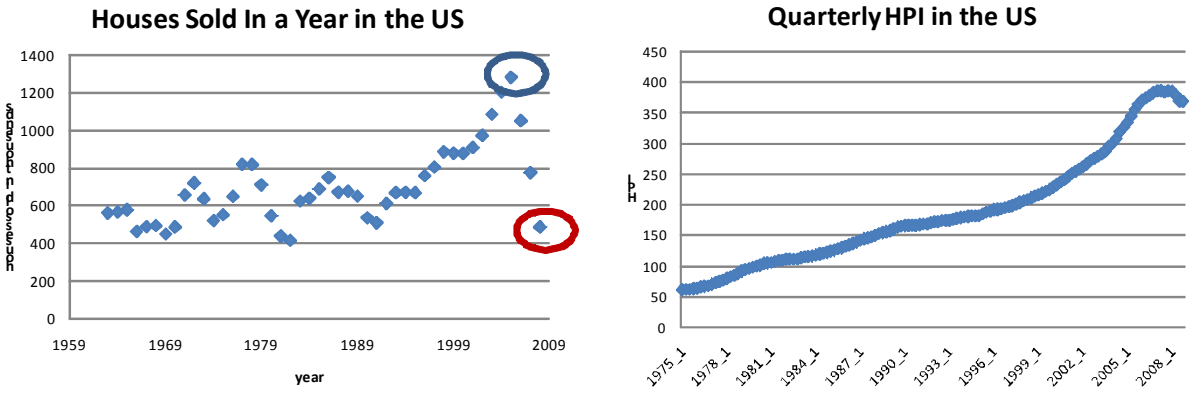

Figure 2: Housing and Prices of New House Sold in the US. (a) Number of New Houses Sold Annually. (b) Quarterly House Price Index.

Google Trend data provides an index for the volume of queries based on geographic locations and time. The search index is a compilation of all Internet queries ever submitted to Google's search engine since 2004. The index for each query term is not the absolute level of queries for a given search term. Instead, it reports a query index measured by query share, which is calculated as the search volume for a specific query in a given geographical location divided by the total number of queries in that region at a given point in time[5]. Thus, the index is always a percentage from 0 to 100. For details on how the query index is calculated, please refer to the Google Trend website at http://www.google.com/insights/search/#.The reports on search index 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 drilled 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 real estate agencies using the frequency of online search terms. From the graph, interests in housing price peaked in 2005 and reached the bottom in 2008.

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

Our analysis uses a predefined category in Google Trend, "Real Estate" to approximate the overall interest for housing. This category aggregates all online search queries that are related to real estate. It further refines the real estate category into several subcategories, such as the "real estate agencies" which collects all search queries related to real estate agencies. Using Google Trend, we generate search index for the general real estate category as well as the real estate agencies category for all 50 states in the US and the District of Columbia from January 2004 to August 2009. We hypothesize that this search index is 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 house price index.

Housing Market Indicators

We collect housing market indicators, such as housing sales and house price index, to examine their relationship with their corresponding online search. The volume of housing sales is collected from National Association of Realtors (http://www.realtor.org/research) for all 50 states in the US and the District of Columbia from the 4th quarter of 2007 to the 2nd quarter of 2009. We also obtain the HPI for the same period from 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 house price index have also increased gradually and reached a peak in 2007, two years after the housing sales peak (Figure 2b). Comparing housing market indicators (Figure 2) to their related online search index (Figure 1) shows that they are closely correlated. As shown in Figure 1, housing related search freqency peaked at 2005 and gradually declined to its lowest point in early 2009, mirroring the volume of houses sold in Figure 2(a). This gives us some evidence that the search index for housing may be correlated with the number of houses sold and has the potential to predict housing sales and the house price index.

Empirical Methods

We use a simple seasonal autoregressive (AR) model to estimate the relationship between the search index and housing market indicators, such as the volume of houses sold and the house price index. A single explanatory variable is used: the search index of housing related queries for each state in the US. We first estimate the baseline AR model and then we add the search index for housing as an explanatory variable to see if it can predict the current financial indicators (sales volume and price index) for the real estate market.

$$HouseMarketIndicators_{it} = \alpha + \beta_1 HouseMarketIndicators_{i,t-1\ t} + \varepsilon \quad (1)$$

$$HouseMarketIndicators_{it} = \alpha + \beta_1 HouseMarketIndicators_{i,t-1\ t} + \beta_2 HouseSearchTerm_{it} + \varepsilon \quad (2)$$

We then add lagged search index to explore if it has any predictive power to forecast economic trends. Since our data is longitudinal, we apply a fixed-effect specification to all our models in order to eliminate influence from any time-invariant properties, such as the demographics of a state, or any state-wide policies that may affect real estate purchase decisions.

$$HouseMarketIndicators_{it} = \alpha + \beta_1 HouseMarketIndicators_{i,t-1\ t} + \beta_2 HouseSearchTerm_{it} + \beta_3 HouseSearchTerm_{it-1} + \varepsilon \quad (3)$$

Next, we examine the predictive power of search index in a supply-demand framework. We use both the volume of houses sold and the house price index simultaneously in a model.

$$HouseSold_{it} = \alpha + \beta_1 HouseSold_{i,t-1} + \beta_2 HouseSearchTerm_{it} + \beta_3 HouseSearchTerm_{i,t-1} + \beta_4 PriceIndex_{it} + \varepsilon \quad (4)$$

Finally, we test if housing related Internet search also spurs future economic activities for industries that complement home-buying activities. For example, if online search can indeed reveal consumers' intention (Moe &

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

Fader, 2004), we may also expect a surge in Internet queries about home appliances, shortly after observing a rise in home sales. Naturally, new home owners may plan to purchase appliances to furnish their properties. By tracking their online search behaviors, we may be able to detect their plan to purchase home appliances. To test this proposition, we correlate housing sales with the search index for home appliances. If the search index for home appliance can translate into actual purchases, we would expect a rise in home sales to indicate future demands for home appliances.

$$HomeApplianceSearch_{it} = \alpha + \beta_1 HouseSold_{it} + \beta_2 HouseSolds_{i,t-1} + \varepsilon \qquad (5)$$

Empirical Results

Predicting Home Sales Using Online Search

Using the search index captured by Google's search engine, we show a positive relationship between housing sales and online queries related to real estate (Table 1). All the models in Table 1 are seasonal autoregressive model, estimated using fixed-effect specifications. The basic assumption for seasonal autoregressive model is that the sales for the current period are related to sales from the previous period. We see a broad support for this as the lagged sales are strongly correlated with the contemporary sales. We apply fixed-effect specifications to all the models in Table 1 since they eliminate influence from any time-invariant characteristics such as the relative size and population of the states, as well as average disposable income of its citizens. We also added seasonality controls for time-specific changes by creating a set of dummy variables for each quarter of the year. To capture online interests in purchasing real estate properties, we use the search index for a predefined category in Google Trend—"Real Estate"—that contains all queries pertaining to real estate. However, we realize this category may be too broad to infer the underlying home buying activities, since it can contain queries that may have minimal relevance to buying or selling a home. Therefore, we also use the search index for the category "Real Estate Agencies" to approximate consumers' interests in real estate. We assume people who are looking for real estate agents may be more likely to buy or sell a house.

Overall, the contemporaneous search index is not shown to have any predictive power on housing sales. As shown in Model 1 of Table 1, we fail to detect any statistically significant relationship between the present search index and housing sales. Interestingly, we find the past search index on the category "Real Estate Agencies" to be positively correlated with home sales. As shown in Model 2 in Table 1, a 1% increase in the search index for the category "Real Estate Agencies" from three months ago is correlated with selling an additional 67,210 new homes in the contemporaneous quarter. This provides some evidence that online search behaviors have the predictive power to forecast future economic activities.

Next, we examine both the current and past search index and their correlations to housing sales in a single model. As shown in Model 3 in Table 1, the past search index about the "Real Estate Agencies" category remains to be positively correlated with the present housing sales ($\beta = 67.70$, $p < .05$) while the current search index fails to be a statistically significant predictor. This further demonstrates that past search behaviors has the power to forecast housing sales, even more than the contemporaneous search activities. In Model 4, we explore the effect of search index on the category "Real Estate". We find the search index for the category "Real Estate" is not correlated with housing sales, as the coefficient estimates on the current and past search index fail to produce any statistically significant results. This could be due to the fact that the "Real Estate" category may contain queries that are not necessarily related to housing sales. For example, queries about home insurance or property managements, which have little relevance to housing sales, are a part of the "Real Estate" search index. However, we continue to find that past search index for the category "Real Estate Agencies" to be positively correlated with housing sales while its present search index is not. In Model 5, we eliminated current HPI from the regression and we find similar results as in Model 4 where HPI is included.

Table 1. Linear Regression on Sales Volume on Search Term Related to Real Estate						
Dependent Var.	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales
Model	Baseline	(1)	(2)	(3)	(4)	(5)
	Fixed Effect	Fixed Effect	Fixed Effect	Fixed Effect	Fixed Effect	Fixed Effect
Quarterly sales from 3 months ago	0.003 (0.107)	0.200** (0.088)	0.219** (0.085)	0.210** (0.087)	0.227** (0.0881)	0.039 (.105)
Current HPI		-1.045*** (.116)	-1.099*** (.108)	-1.069*** (0.116)	-1.027*** (0.119)	
Current Search Index on "Real Estate Agencies"		23.42 (33.09)		24.81 (32.70)	30.90 (43.47)	125.12* (61.79)
Past Search Index on "Real Estate Agencies"—3 months ago			67.21** (16.84)	67.70** (31.64)	121.4** (46.85)	163.44*** (57.349)
Current Search Index on "Real Estate"					-25.77 (60.77)	-41.06 (74.76)
Past Search Index on "Real Estate" from 3 months ago					-89.52 (58.02)	-204.11 (169.52)
Obs.	304	304	304	304	304	304
Controls	Quarters	Quarters	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51	51	51
Adjusted R^2	0.973	.982	.982	.982	.984	.980
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

Predicting the House Price Index Using Online Search Data

In Table 2, we explore the relationship between the housing related search index 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 2 evaluates a seasonal autoregressive model using fixed-effect specifications. As predicted, the lagged HPI is positively correlated with the present HPI. In Model 1, we estimate if the current search index on the category "Real Estate" and "Real Estate Agencies" have any predictive power on the current HPI. Interestingly, both current and past search index are positively correlated with the contemporaneous HPI. As shown in Model 1, a one percent increase in the current search index for the category "Real Estate" is associated with an increase of 29.57 points in the current HPI. Similarly, a one percent increase in the search index for "Real Estate Agencies" is correlated with an additional 10.12 points in HPI.

In Model 2, we introduce both the current and the past search index in the model. Interestingly, the search index for the category "Real Estate Agencies" are no longer shown to have any predictive power for forecasting the current HPI while the coefficients for the search index for the category "Real Estate" becomes statistically significant. We find that the current search index for the category "Real Estate" remains to be positively correlated with the current HPI. This appears to reflect an increase in housing demand, driven by home buyers who search for properties online prior to actually buying. We find that a one percentage increase in the search index on the category "Real Estate" is associated with 37.83 points or a 10% increase in HPI. As buyers are more likely to conduct search online prior to buying, past search index can be used to predict future housing price.

Table 2. Linear Regression: Correlations between HPI And Search Index		
Dependent Var. (monthly, in Thousands)	HPI by State	HPI by State
Model	(1)	(2)
	Fixed effect	Fixed effect
HPI(t-1)	1.062*** (0.0172)	1.054*** (0.0178)
Real Estate Agent Search Index—current	10.12** (4.917)	-8.061 (6.173)
Real Estate Agent Search Index—3 months ago		27.04 (16.243)
Real Estate Search Index — current	29.57*** (8.814)	37.83*** (8.752)
Real Estate Search Index –3 months ago		-22.27*** (8.504)
Controls	Quarter	Quarter
Obs.	408	357
States	51	51
*p<.1, **p<.05, ***p<.001, Huber-White robust standard errors are shown in parentheses		

Interestingly, we find that the correlation between the HPI and the search index for the "Real Estate" category is negative if the search phrase happened three months ago. We infer this to correspond to an increase in the supply of available houses in the market, shifting the supply curve outward instead of the demand curve for buying houses. We find that a one percent increase in search index is associated with a drop of 22.27 points in the HPI. Sellers "move first" in this marketplace, surveying the competition and assessing market conditions before making a decision to sell. As more sellers reveal their intentions, more houses eventually become available for sale, shifting the supply curve outward. In turn, the listing price is likely to go down and decrease the overall HPI. Consequently, we observe that the current online search on housing begin to drive down the HPI three months after the search.

Figure 3 provides a rough sketch of the housing demand and supply curves. As we can see in the figure, an increase in the search index from 6 months ago shifts the housing supply outward, as indicated by the downward arrow (a). When the housing supply increases, housing sales are likely to rise, moving the quantity of houses on the market from Q1 to Q2 where Q2> Q1. These search terms, conducted 6 months ago, reflect sellers' presale activities as they assess the market condition and surveying the competition. As more sellers reveal their intentions and more houses become available on the market, the overall housing price as measured by house price index goes down as well (P2<P1).

On the other hand, current search terms may reflect buyers' pre-buying activities as they search for properties online. These properties are likely from sellers who had previously explored the housing market online 3 months earlier. As buyers' interests grow, the demand curve shifts outward (arrow (b) in Figure 3)), driving both the price and quantity of houses upward (Q3 > Q2, P3 > P2).

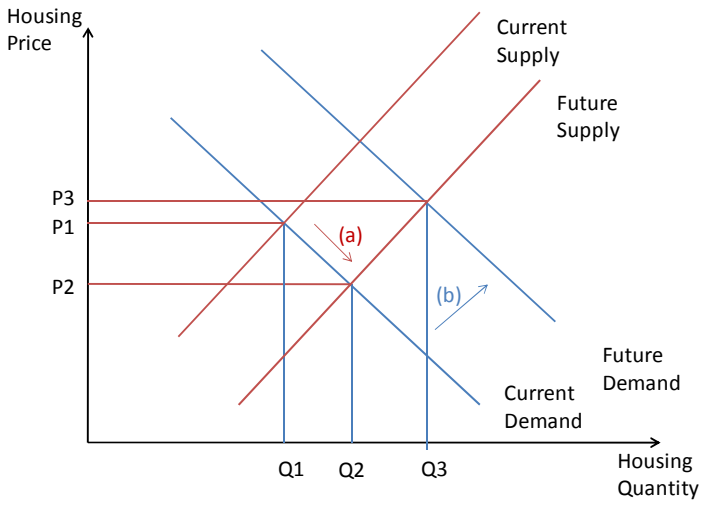

Figure 3: Demand-Supply Curve for Housing

Predicting the Demand for Home Appliances

Lastly, we explore trends in sales of home appliances. We expect that housing sales would spur interests in home appliances, and thus increasing their demand in the future. To gauge the overall interests in home appliance, we use the search index for the home appliance category from Google Trend. In Table 3, we show its relationship with home sales. We observe that the current home sales are not correlated with the present interest in appliances (Model 1, Model 4). But after 6 months, each 1000 houses sold previously is correlated with a 1.14% increase in the search index for home appliances. Since buyers move into their new property first before they start making any major purchases, it is natural that interests in home appliances would increase only after a consumer has already bought a house. Thus, if the current search index for home appliances is predictive of their future demand, we may expect online search about home appliances, spurred from housing sales, to indicate future purchases of home appliances. This highlights the linkages between home sales and other parts of the economy that may complement home sales.

Table 3. 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)
Model	(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 trillions 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 not reserved to an elite priesthood, but are increasingly available to ordinary consumers, business people and researchers of all types. While much of social science research in the past several decades has focused on ever more sophisticated statistical and mathematical models, for many types of important business and economic predictions, IT research is slicing the Gordian knot by making available – and putting to work – a several orders of magnitude improvement in data quality.

In particular, we have shown that analyzing online search data with relatively simple models can yield surprisingly accurate predictions about the housing market. 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 the housing search index is correlated with both housing sales and the house price index. This correlation lends support to the hypothesis that Web search can be used to predict present and future economic activity. For example, housing-related search can be used to predict the recovery of the currently embattled housing market and potentially, when the economy may recover from the current recession.

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 home buyers 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 Trend 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.

In addition to mattering for consumers and businesses, benefits exist even at the level of the whole economy. By decreasing the information disparity between consumers and the market, the use Google Trends data that we describe can also improve the efficiency of the housing market, saving billions of dollars and freeing valuable resources for long term growth. 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. We find that housing sales are also positively correlated with future search frequencies on home appliances. If search frequencies predict future demand, we would also expect an increase in demand for home appliances after a surge of interest from house-related queries. Thus, observing an increase in housing related searches can help the home appliance industry to adjust its future inventory accordingly.

Other applications of this research

While we show promising predictions about the housing market using Google Trend, 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 between HD DVD and Blue-Ray. If user interests for Blue Ray grow overtime relatively to HD DVD, we may expect Blue Ray to win the standards war. Similarly, we can also 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. As a simple example of this, we show that online search for Apple laptops are correlated with sales.

As shown in Figure 4, we see a dip in user interests in Apple laptops in March, 2006 and at the same time, sales of Apple laptops also dropped. Similarly, user interests for Apple laptops reached a peak in September, 2008, and the sales for Apple laptops in the 4th quarter of 2008 also rose to its highest value since 2004, suggesting that the current online search can forecast sales. Although this analysis is at a very preliminary stage, we demonstrate that online search can be used to model future IT product sales and market shares in a timelier manner than quarterly industrial reports. As with the housing market, analyses of IT products and services can be made more robust by collecting additional data and control for factors that may influence sales, such as the overall interests for electronic goods.

Thanks to ubiquitous access to the Internet and search engines among consumers, we are now able to capture consumer behavior at a level that has not been achieved before. For instance, trending data collected from Google Search enables us to track consumer interests in several major metropolitan areas of the world on a weekly basis. IT research, however, does not appear to be at the leading edge of this thriving field of gleaning collective knowledge from micro-level data points.

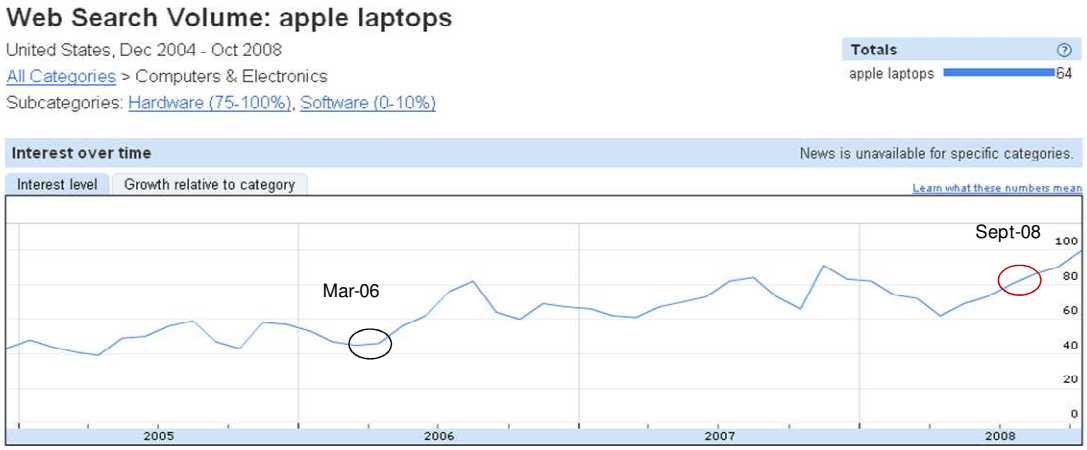

Figure 4a: Google search frequencies for Apple laptops.

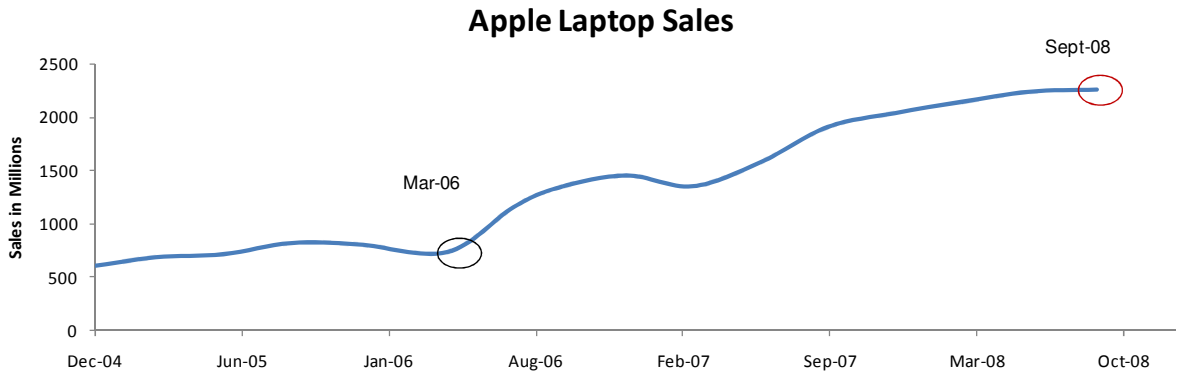

Figure 4b: Actual sales of apple laptops from Apple's 10Q reports.

Conclusions, Limitations and Future Work

Almost exactly 400 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, vaccines, medicines and treatments that saved innumerable lives. 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 housing sales and the house price index, lending credibility to the hypotheses that Web search can be used to predict future economic activities, for example, when the economy may recover from the current recession. We are aware of the fact Google search queries do not represent all the online activities related to searching for a house. 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. Using data from Google Search alone would miss this type of consumers. However, despite missing data, we can still accurately predict the housing sales and the house price index using online search data captured by Google. This demonstrates the power of using online queries to forecast economic trends.

To extend our results, we can collect additional housing related data, such as interest rates, policy changes, and the unemployment rate in each state over time. However, using a fixed-effect model and controlling for seasonal effects provide us some confidence in our results. Fixed effect model can eliminate influence from time-invariant properties and seasonality can control for time-specific changes, such as the fact that houses are less likely to be sold before the holidays. We would also like to perform more fine-grained analyses at the MSA (Metropolitan Statistical Area) level. Although state-level analyses are useful, analyses at the MSA level would be more beneficial as housing market trends are more dependent on MSAs.

The abundance data from Google Trends allows researchers to study numerous topics. Currently, we are also in the process of studying technology related trends using Google Trend. For example, we can use search index to predict if Macintosh computer will gain market share against Windows operating systems. Similarly, we can explore the market shares for various portable devices such as cellular phones and MP3 players.

Ultimately, micro data collected using Google Trend may prove one of the most powerful tools that IT research has delivered 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. Nonetheless, as these data 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.

Reference

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

Hyunyoung Choi, Hal Varian, 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

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

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", 2009

Nielsen Report, 2008, http://www.polepositionmarketing.com/emp/august-2008-search-3/

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

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.