Big Data in Practice

Through our conversations with clients, the most common questions that come up related to Big Data are: "I know I need to do something about Big Data but where do I start?" and/or "Where can I find these new data sources?". There are potentially hundreds if not thousands of vendors who are part of this Big Data ecosystem but it is hard to differentiate the most alpha-adding datasets without investing substantial amount of time in researching them in order to have a priority list. In this report, we attempt to help investors answer both questions by investigating different types of Big Data sources and what their potential uses could be.

New Datasets = Alternative Data

To be clear, the 170+ vendors (including government and other free data sources) we have listed in the following sections and the appendix by no means constitute the complete set of Big Data vendors as more companies are being set up all the time. These have been included as representative examples in each category to highlight what's possible and potentially useful. We discuss two types of vendor solutions: one is data directory and analytics services that help investors locate unusual datasets and offer advice in their usage, whilst the other covers analytics providers who offer specific types of data that are semi-structured, derived from unstructured datasets.

Eagle Alpha

Through our research process, amongst other vendors who specialize in one or two types of big data, we discovered an alternative data and analytics provider called Eagle Alpha. In 2012 Emmett Kilduff, a former Morgan Stanley investment banker, set up Eagle Alpha with the basic premise that the amount of alternative data being created worldwide was growing rapidly but only a handful of asset managers were able to take advantage of it. The aim of the company is to be the 'one-stop shop', for asset managers who want to explore alpha opportunities from alternative data.

Eagle Alpha currently has a team of twenty employees and nine advisors. Senior employees have substantial experience in alternative data and investing with one member being a former Director of Equity Research at a large US asset management firm. In addition, their advisory panel consists of a former head of R&D of a Big Data unit at a large quantitative US hedge fund. Their broader team includes data procurement professionals, research analysts, data scientists and engineers.

Eagle Alpha's product offering has evolved over the last five years to reflect the needs of its clients and in turn the maturing of Big Data usage. There are six parts to its offering – see Figure 8 overleaf.

Figure 8. Eagle Alpha Overview

Source: Eagle Alpha, Citi Research

Specifically on data sourcing, Eagle Alpha's Data Sourcing team is experienced at sourcing, appraising and performing due diligence on companies which possess alternative data. Their database consists of almost 500 data sources as of Feb 2017 mapped to their taxonomy of alternative data categories, GICS, MSCI market classification and eight macro categories[12].

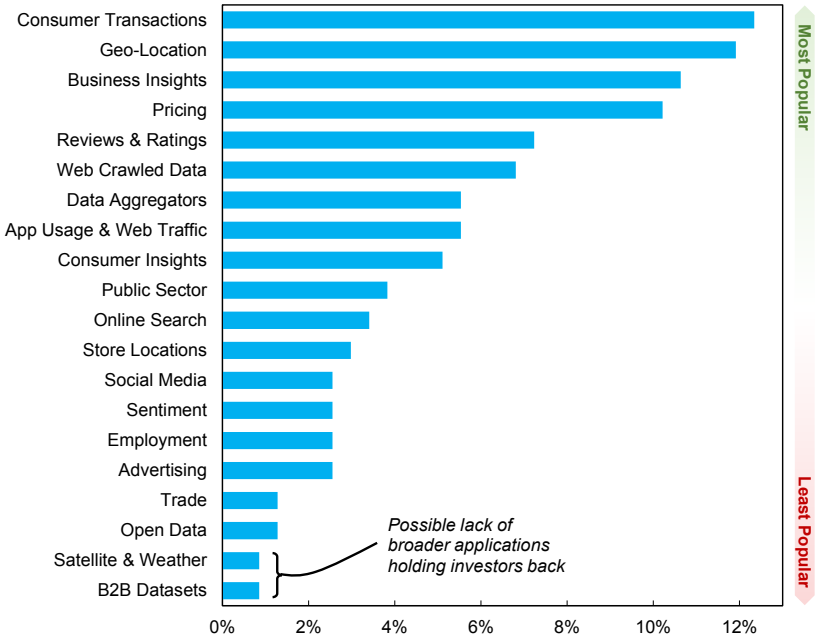
As part of their offering, Eagle Alpha provides customers with a web platform which lists a large number of alternative datasets. This platform enables users to see brief descriptions of datasets/data vendors, historical data availability, API access and whether it's available freely or of paid access. In addition, the company enables all data vendors to complete self-populating profiles on the Eagle Alpha platform which helps their customers prioritise the datasets.

On top of being a data provider, Eagle Alpha offer advisory services where scheduled calls are arranged with their Head of Data Sourcing on the latest datasets and advise how best they can be used to address their research questions and investment debates. Finally, they host exclusive data showcase events where selective owners of alternative data present their offerings.

[12] For information about their data sourcing process, please contact Emmett Kilduff, CEO, emmett.kilduff@eaglealpha.com, Eagle Alpha

Based on their taxonomy of data categories, Figure 9 shows the most interest they have received surrounding consumer transactions and geo-location data, whilst satellite and weather datasets are the least popular. This may be an indication of the complexity in analyzing satellite datasets and the lack of broader application of this data that is restricting investors interest.

Figure 9. Alternative Data Interest by Category



Source: Eagle Alpha

In the following section, we apply some of the selected datasets to see whether or not the new alternative information could add value to investors or their process.

Case 1: Google Trends for US Unemployment Prediction

Starting with macro-related alternative data, we investigate the appeal of web-search data to predict macroeconomic indicators which in academic literature shows some promising results.

Google Trends is a publically available web facility based on Google Search that shows how often a particular search-term is entered relative to the total search-volume over time across various regions of the world. There are several advantages to using search data as a determinant of economic activity – (1) the data is more timely than traditional datasets as it's available in a higher frequency; (2) it has over ten years of history; and, (3) it offers flexibility in terms of a variety of investment questions that can be analysed. The data is also generated as a by-product of people's normal day-to-day activities (exhaust data), as opposed to traditional survey methods which rely on individuals or firms responding to survey questions after the event. This can avoid problems associated with non-response or biased survey results.

However, it is not so easy to come up with a 'bullet-proof' ontology or a complete list of relevant search terms for any analysis one wishes to conduct. For example, how could one distinguish whether the search was related to apple, the fruit, or Apple,

the tech giant? It requires a thorough think-through and collaboration with data scientists and advanced machine learning to achieve more reliable outcomes which take time and experience to perfect. One provider on Eagle Alpha's platform has historical data for up to 120 million search keywords across 27 countries.

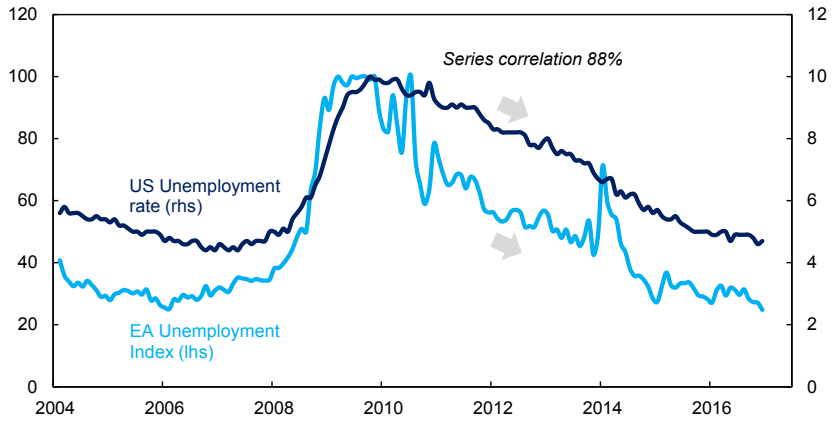
We understand that Eagle Alpha (EA) data scientists and analysts have spent almost three years finding the best way to use online search data to predict economic indicators. They have devised their own proprietary methodology that leverages relevant academic research, as well as accepted best practices in the field.

Each index they have built based on Google Trends follows the below process: (1) generate relevant search terms; (2) source the search volume for each term dating back to 2004; (3) curate the data and adjust for outliers and seasonality; (4) search terms are ranked by their predictive scores; and, (5) final index includes a selected basket of terms, and measures co-movement of search activity with a particular economic indicator.

The aim of these indices is not to provide point estimates for macro investors, but instead designed to capture inflection points more accurately and thereby improve the predictive power of investors' estimation models.

We have obtained EA's US unemployment index data, which is constructed to measure online search activities relating to claims of unemployment benefits. The data goes back to 2004 and compares the trends/directions of this index against the official released unemployment rate from US Unemployment Bureau of Labour Statistics. Figure 10 shows the historical movements of these two data series with the correlation between the two series being 0.88.

Figure 10. US Unemployment Comparison



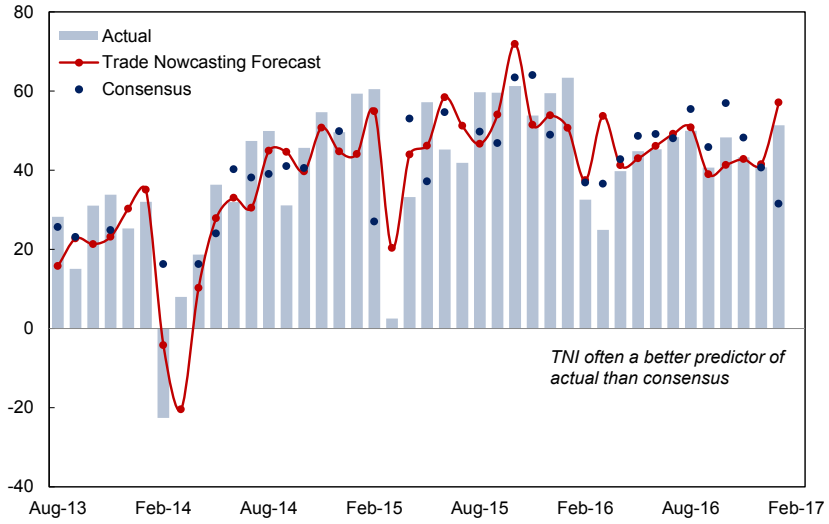
Source: Eagle Alpha, US Unemployment Bureau of Labor Statistics, Citi Research

Compared to a simple model of using 12-month moving average as the predictor for changes in unemployment, the EA unemployment index has a higher correlation of 51% with the subsequent labour market movements, compared to 46% based on the historical reported numbers. Having that said, we do not envisage this index being used on a standalone basis as economists and asset allocators typically have a wealth of other inputs in their toolbox that help them to form their predictions. As far as the EA unemployment index is concerned, whilst prone to some criticism on the assumption that unemployed people all have access to internet, this case study however does demonstrate the potential value-add to the decision-making process from an on-the-ground assessment angle.

Case 2: Trade Nowcasting for China's Trade Balance

Another interesting dataset we've come across within EA's offerings is Trade Nowcasting Indicators (TNI). TNI provides valuable insights into international trade and industrial production through Big Data and predictive analytics techniques. Coupled with more than 25,000 time series, they are able to forecast trade balance and industrial production statistics. The data is also available at the level of an individual shipping port, providing even greater granularity on international trade.

Figure 11. China Trade Balance – Reported vs Forecast

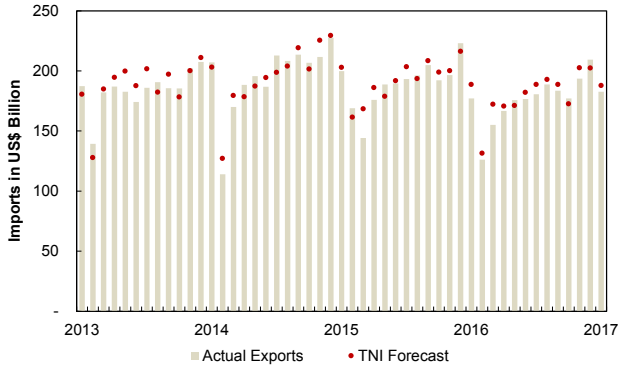


Source: Eagle Alpha, Citi Research, Bloomberg

As China remains a country of interest for many investors, we have obtained the indicator for China's trade balance from EA's data partner and have compared the forecast to the actual reported figures, and also consensus. Note that in the early part of sample, the consensus was quite sparse with intermittent data being available. We have plotted the series in Figure 11 from the start for completeness but we would recommend focusing more on the recent observations where there is a more complete representation from street analysts/economists. From the chart, it is apparent that the TNI forecast is often a better predictor of the actual number, compared to consensus. In the last 12 months, close to 70% of the time, TNI has predicted the trade balance closer to the reported figure than consensus, notwithstanding the advantage of substantial lead time over the official data release.

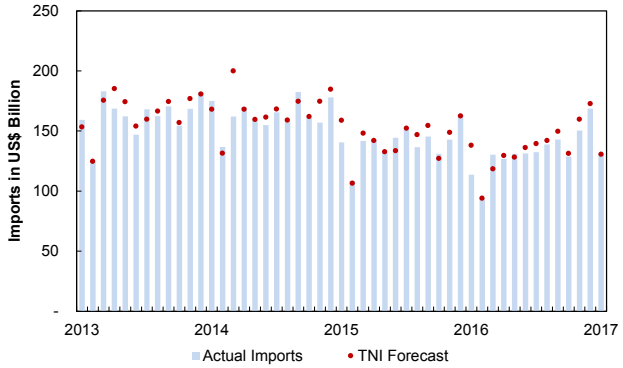
Another interesting aspect of this data is that it provides both exports and imports forecast for China. Figure 12 and Figure 13 demonstrate that on a more granular level, TNI forecast appears to capture the inflection points well on both sides with correlation between actual change and forecast change of exports and imports reaching 88% and 85%, respectively. This could be a very valuable for asset allocators and top-down investors as they watch the direction and magnitude of the changes closely to form their investment decisions.

Figure 12. China Exports – Actual vs TNI Forecast



Source: Eagle Alpha, Citi Research, Bloomberg

Figure 13. China Imports – Actual vs TNI Forecast



Source: Eagle Alpha, Citi Research, Bloomberg

Turning our attention to stock selection ideas, we investigate three company cases where alternative datasets are used to predict individual stocks' sales/revenue growth.

Case 3: Google Trends for Burberry Same Store Sales Forecast

As mentioned previously, Google Trends is a public web facility based on Google Search. Using Google Trends, EA has also built company specific indices based on search terms that are related to a given retailer's product offerings. This constitutes an exhaustive process within EA for identifying search terms related to a company's revenues using both internal and third party tools. In the case of Burberry, we understand EA has deployed over 20 related terms in order to comprehensively capture consumer interests in Burberry's products.

Before we proceed with the test, we have the prior that for searches/Google Trends data, there might be information when shorter-term measures 'cross-over' longer-term averages. This is intuitive as consumers might be drawn to a new collection or celebrity endorsement which contributes to a sudden increase of interest that is likely to translate into sales. As Burberry reports its same store sales numbers on a quarterly basis, we believe one-month data points crossing over three-month moving averages of these search indices might have predictive power of inflection points in revenue growth.

Figure 14 appears to support our prior that short-term 1-month YoY observation crossing over the 3-month moving average YoY indicates major inflection points of same store sales growth as depicted by the green circles on the chart. Another interesting finding is that consensus[13] is pretty bad at predicting same store sales growth. In fact it only achieves close to 20% of correlation with actual reported figures. With either the 1-month YoY or 3-month moving average YoY measures based on EA equity index for Burberry, the correlation jumps to over 70% which is a significant improvement. The additional advantage of this search data is its timeliness – the data at the end of the quarter is available immediately, whilst official figures typically are announced at least 3 weeks after quarter-ends. The timeliness and much improved correlation of the new dataset with actual reported figures make such an offering appealing.

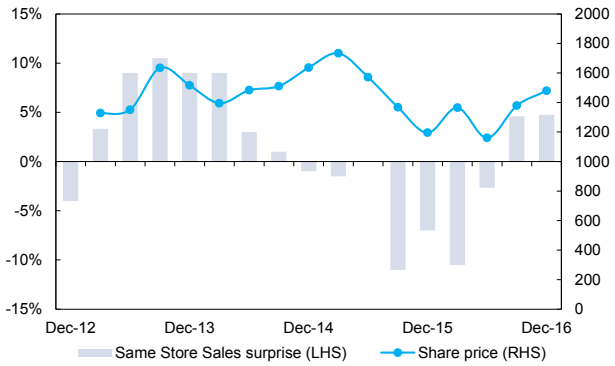
[13] The consensus data is sourced from Bloomberg but it has low analyst coverage issues as not all analysts publish their estimates for same store sales.

In Figure 15 we highlight that same store sales does have pricing impact especially the surprise element of it. That is, markets react to the positive/negative sales surprise. This suggests that, if we are able to predict the inflection points better with Google Trends data, there could be pricing implications from being able to act sooner and more accurately than the bulk of investors.

Figure 14. Burberry Same Store Sales vs Eagle Alpha Stock Index (green circles denote significant cross-over of 1M vs 3M indicating potential inflection points)

Source: Eagle Alpha, Citi Research

Figure 15. Burberry's Same Store Sales Surprise vs Share Price



Source: Bloomberg, Citi Research

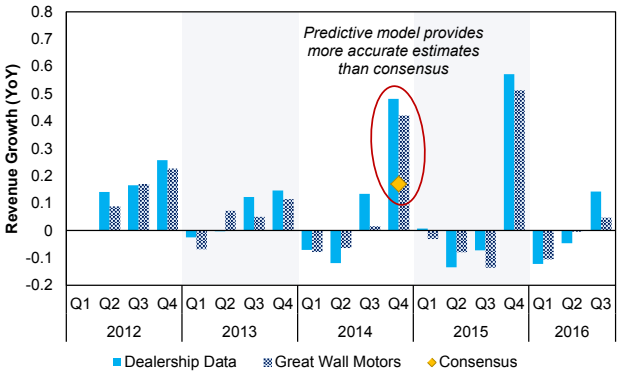
Case 4: Dealership Data for Great Wall Motors Sales Forecast

According to Eagle Alpha, their China Auto Insight (CAI) dataset is provided through an exclusive partnership with a leading Chinese financial automotive consultant. The dataset is collected using a large panel of dealerships throughout China, as well as combining other data sources such as web data and more traditional data sets to create a large and well-structured database.

Using the CAI dataset, Eagle Alpha have calculated dealership revenues for Great Wall Motors (GWM) from the beginning of 2012 and they have modeled company revenue using dealership revenue as a predictor. The dealership revenue displays a 99% correlation with reported revenue for the company, and a 95% correlation with YoY growth rates in reported revenues. The calculated dealership revenue also correctly projected the directional movement of reported revenues for GWM in fourteen of the fifteen quarters between Q1 2013 and Q3 2016. In addition, according to Eagle Alpha, the prediction model they have built shows that its measure of dealership revenue is a strong predictor of GWM's revenues with an error rate of just 4.9%, as depicted in Figure 17.

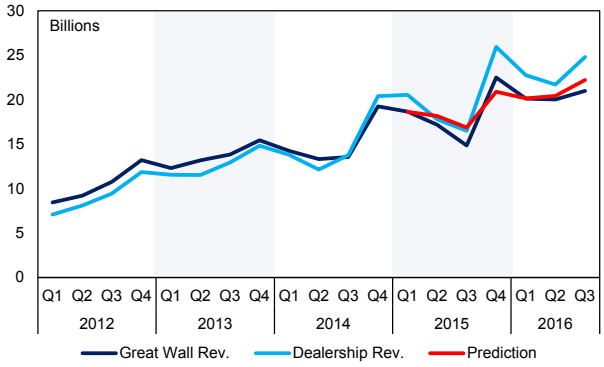
The result is intuitive as dealerships are the frontline of GWM sales and if one could have access to such data, there is no question that the aggregated data would have high accuracy in predicting the eventual reported sales. The information edge over the street estimates is potentially large. For example, as shown in Figure 16, in Q4 2014 the CAI data was forecasting a QoQ growth rate of 48% compared to the consensus estimate of 17% versus the reported QoQ growth rate of 42%

Figure 16. Great Wall Motors Revenue Growth vs CAI Dealership Data



Source: Eagle Alpha, CAI Data, Bloomberg, Citi Research

Figure 17. Great Wall Motors Revenue Prediction



Source: Eagle Alpha, CAI Data

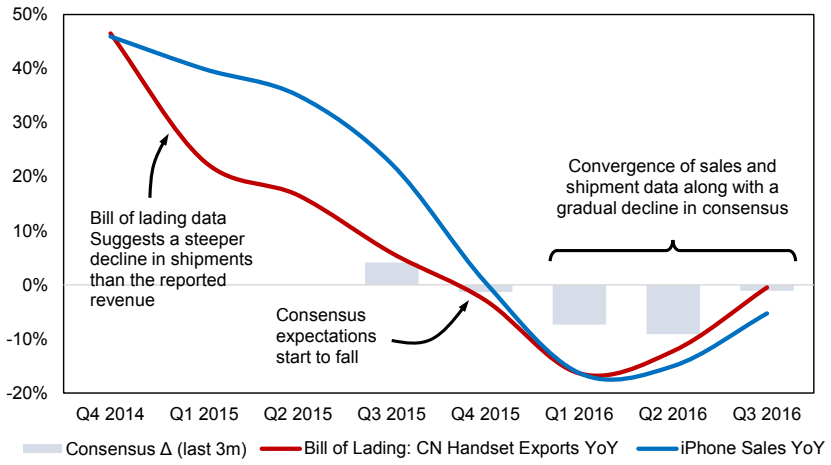
The analyst estimates in this calculation were taken one month after the end of the quarter, whilst the CAI data is published between five and twenty days after month-end. It is evident that the predictive model based on the CAI data can provide investors with a more accurate revenue estimate for Great Wall Motors.

Case 5: Bill of Lading Data for Apple iPhone Sales Forecast

A bill of lading is a detailed list of a ship's cargo in the form of a receipt given by the master of the ship to the person consigning the goods. Bill of lading data from multiple countries can be aggregated and structured at the company-level to provide insights on companies engaged in global trade. This data provides exceptionally granular insight into shipping trends in a timely fashion. The obvious applications of such data include forecasting economic variables such as trade balances and also individual company revenues. Eagle Alpha's bill of lading data partner aggregates and structures company-level import and export data from multiple countries to provide insights on global trade activities.

As Apple's iPhone is known to be produced in China, we have obtained the bill of lading data from EA to examine whether such a dataset could predict iPhone sales in a timely manner. Figure 18 shows that the bill of lading data suggested a steeper decline in iPhone shipments in 2015 than was actually reported in iPhone sales. This was down to iPhones produced in 2014 still being sold through the Apple sales channels, masking the impending weakness of sales growth as indicated by the shipment data. It was not until the fourth quarter of 2015 that expectations for Apple iPhone sales began to decline. The convergence of iPhone sales and shipment data began in the first quarter of 2016 and went on for three quarters. The 'over-shipment' in 2015 lasted approximately three quarters and the adjustment of expectations also took three quarters as consensus showed a gradual decline.

Figure 18. Bill of Lading versus iPhone Sales versus Consensus



Source: Eagle Alpha, Bill of Lading Data, Bloomberg

Apple started to lower its guidance in Q1 2016 which led to a decline in consensus forecasts in the quarters of March and June 2016. In the chart above, the grey bars represent the change to consensus estimates three months prior to Apple reporting. In other words, this is the change to consensus due to management revision of forward revenue guidance. With bill of lading data, we are able to uncover revenue growth as reported by the company and spot the weakness in demand at an early stage, months before the reported decline and lower guidance provided by Apple management which resulted in a significant pricing impact.

This case study illustrates that new datasets do not only provide potential short-term advantages, but they could also have longer-term applications for fundamental stock pickers who have longer investment horizons.