Deutsche Bank
Markets Research

North America
United States

Quantitative Strategy
Signal Processing

Date
17 February 2016

Big Data in Investment Management

Gaurav Rohal, CFA
gaurav.rohal@db.com

Javed Jussa
javed.jussa@db.com

Yin Luo, CFA
yin.luo@db.com

Sheng Wang
sheng.wang@db.com

George Zhao
zheyin.zhao@db.com

Miguel-A Alvarez
miguel-a.alvarez@db.com

Allen Wang
allen-y.wang@db.com

David Elledge
david.elledge@db.com

North America: +1 212 250 8983
Europe: +44 20 754 71684
Asia: +852 2203 6990

The Big Data Renaissance

Big data is big

Did you know that over 100 million websites are added to the Internet each year? Currently, there exists over one billion websites which represent over 10 trillion individual web pages. Google has indexed more than 200 TB of data. This may seem sizeable. However, it represents less than 0.01% of the 500 Exabytes of accessible data on the Internet. To refresh your memory, an Exabyte is one to the power of 18 bytes or one million terabytes; or one billion gigabytes. It's a lot!

Harnessing Big data

In this paper, we discuss the opportunities and challenges of adopting Big data in investment management. We analyze a wide spectrum of Big datasets from satellite imagery, web mining, social media, textual, crowd sourcing to accounting, macroeconomic, and even IRS tax filings. Additionally, we discuss various analytical frameworks for analyzing Big data such as machine learning, deep learning, and graph theory. Lastly, we outline the key infrastructure elements needed to integrate Big data such as programming languages (e.g., R and Python), cloud computing (e.g., Amazon Web Services), distributed file systems (e.g., Hadoop); the list goes on. Sit back, get comfortable, and enjoy your tour into the exciting world of Big data.

Source: gettyimages.com

Table Of Contents

A letter to our readers

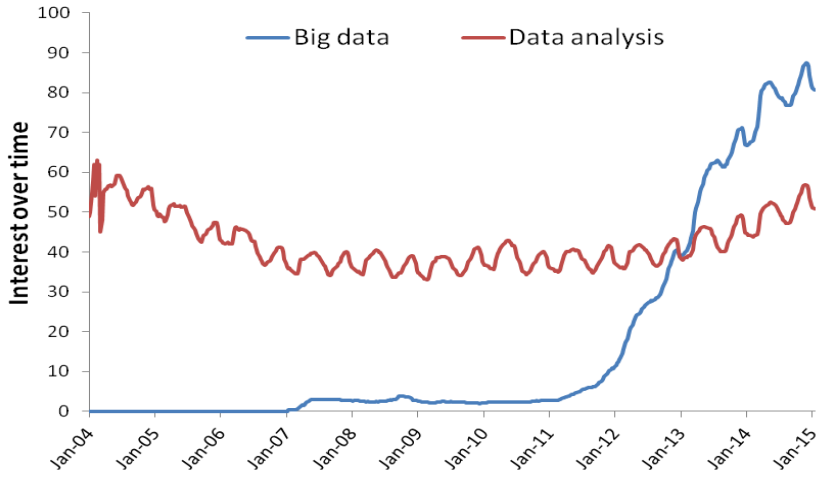
The amount of data that the world collects has experienced explosive growth. The rapid development of informative technologies has paved the way for a transition from analog to digital. This evolution is further exasperated by reduced technology costs and rapid software development cycles.

The rapid advancement in the Internet and social media sectors has far reaching effects not only on our day to day lives; but also, for investment managers. Analyzing Big Data has become important yet challenging. Ever evolving tools and skills are required to process, store and analyze large amounts of data from a multitude of sources.

A wide array of information sensing devices has led to this explosive growth in big datasets. Starting from handheld mobiles, cameras, microphones, smart cards, wifi linked products, to RFID and satellite imaging

So what is Big Data? A wide array of information sensing devices has led to this tremendous growth in big datasets. Starting from handheld mobiles, cameras, microphones, smart cards, WiFi linked products, to RFID and satellite imaging. Figure 1 shows the interest level in "Big data" and "Data analysis". This information is from Google Trends[1] and is based on search volumes. Clearly the interest level in Big data has intensified.

Figure 1: Over time interest in search terms, Big Data and Data analysis



Source: Google, Deutsche Bank Quantitative Strategy

According to Gartner research, "Big data is high-volume, high-velocity and/or high-variety informational assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation".[2] Overtime new terms have been added to this definition such as variability, veracity, complexity etc.

The sheer magnitude and complexity of Big data makes it difficult for investment managers to use, process, and understand. However, we strongly believe that portfolio managers who invest the time, resources, and energy into Big data and modeling will reap the benefits.

[1] See our previous research in Luo et al. [2014] on how to use Google trends in investment
[2] For more information, see http://www.gartner.com/it-glossary/big-data/.

The volume and complexity of data investment managers have to deal with is only going to increase in future, and we believe there might be alpha on the table for those who can best manage and manipulate these datasets. Better technologies can be a source of alpha in its own right – it may provide the extra edge for portfolio managers. Please contact us at DBEQS.Americas@db.com for more information on Big data as well as modeling techniques. We hope you enjoy the remainder of this report.

Yin, Javed, Gaurav, Sheng, and the quant team

Deutsche Bank Quantitative Strategy Team

Big data analytical framework

What is your problem?

Before outlining the various datasets classified as Big data, we briefly discuss the various data modeling techniques that can be employed. There is an abundance of Big data analytical approaches. Many methods can be employed to solve a particular problem. Some models are better at solving certain types of problems. Irrespective of the method you choose, the problem that you are attempting to solve must be clearly defined. So let's take the time to define the problem that we are trying to solve.

Big data analytical methods can apply to problems within various fields such as the social sciences, biomedical sciences, archeology, genetics, finance, and social media marketing. It is probably no surprise that we choose to employ these modeling techniques within the financial industry; in particular, for stock selection.

Big data analytical methods can apply to problems within various fields such as the social sciences, biomedical sciences, archeology, genetics, finance, and social media marketing

The problem that we lay out is fairly straightforward. We have an ambitious yet practical goal of predicting stock returns or risk. We can design our model either as a classification problem, i.e., which stocks will outperform and which stocks will underperform; or we can directly attempt to get a point estimate of future returns. In order to solve this problem, we need information on the characteristics of these stocks. Naturally, the first place we look for information on stocks are the audited financial statements. We can obtain a wealth of company fundamental information from the financial reports.

Now, we can refine our problem: can we predict which companies will outperform and which will underperform based on a company's fundamental data (or factors) as our inputs. Thankfully, many data vendors provide data on the current and historical fundamental information for most global stocks. Obviously, there is far more data and information about a company that we can use to predict its future returns then merely its fundamentals.

This is in fact a classification problem in statistical jargon. Can we classify which stocks will outperform versus underperform based on the input data. In order for an algorithm to classify stocks as outperformers versus underperformers, it needs to determine what characteristics or factors are telling or predictive of performance. A model essentially learns which factors are predictive of outperformance. This is called model training. The more data and history the model can employ for training purposes, the more accurate are the classification results. Note that our classification problem is designed as a binary problem (i.e. outperformers versus underperformers). This is merely one analytical method of prediction. There are various other analytical methods such as predicting the specific ranking of the output rather than a binary outcome.

Machine learning models

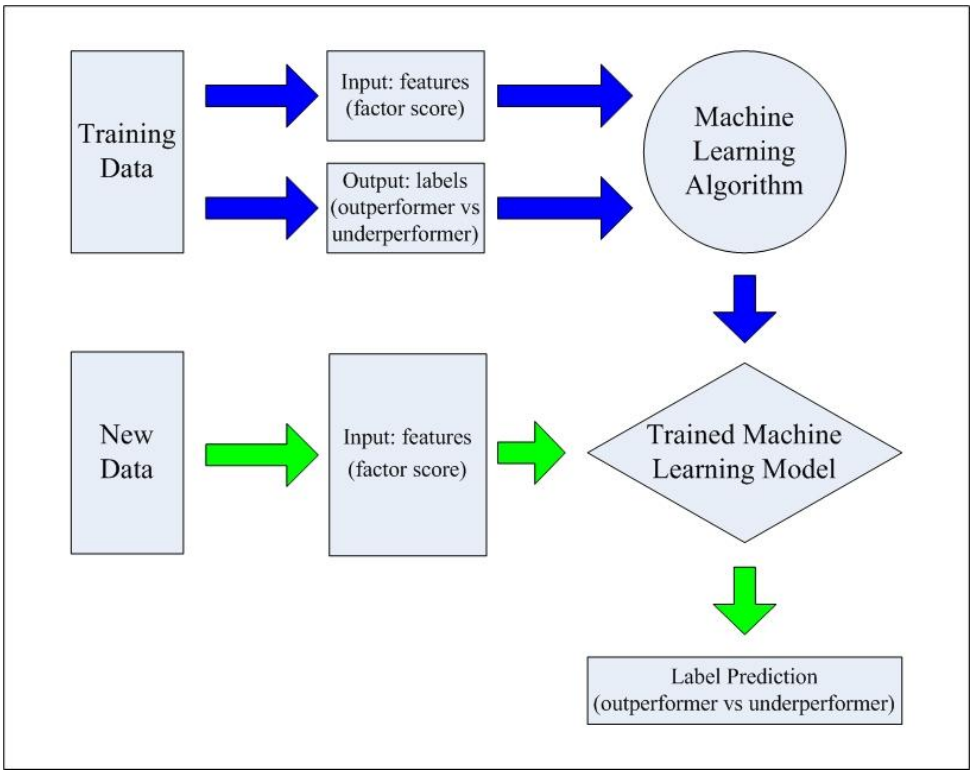
Machine learning is a branch of computer science that is concerned with the design and development of algorithms that allow computers to perform tasks associated with artificial intelligence (AI), based on empirical or historical data. Such tasks involve recognition, diagnosis, prediction, system control, etc. A powerful machine learning algorithm can capture the inherent characteristics of disparate data and identify hidden relationships among observed variables. A major focus of machine learning is to automatically learn to recognize complex patterns and make intelligent decisions based on this analysis. We are all familiar with the Deep Blue computer built by IBM that beat the best chess player Gary Kasparov in 1997. However, it was once believed that a computer probably would take many more years, if ever; to beat a human professional Go[3] player. Therefore, it was shocking that recently Google's DeepMind team built an artificial intelligence system based on "Deep learning" that beat the three time European Go champion Fan Hui.

There are two kinds of learning algorithms unsupervised and supervised models. Supervised learning is often used in the application of prediction. This type of learning trains and labels the training samples based on the actual result, which in turn gives feedback about how the learning is progressing. Supervised learning usually consists of a set of training examples. Each training example is a pair consisting of an input feature and a desired output value or label. A supervised learning algorithm analyzes the training data and produces an inferring function between the input feature and the output value. The inferring function should predict the correct output value for any valid input feature. When it comes time to predict outcomes based on out-of-sample data, the trained learning model will output the label prediction. An illustration of supervised learning is shown in Figure 2.

Supervised learning is often used in the application of prediction. This type of learning trains and labels the training samples based on the actual result, which in turn gives feedback about how the learning is progressing

[3] Go is a popular Asian board game and far more complex than chess.

Figure 2: Diagram of general supervised learning



Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, LinkUp, Deutsche Bank Quantitative Strategy

CART Model

CART stands for classification and regression tree. It is a simple machine learning technique that seeks to classify data into binary branches or trees, by applying hierarchical splits based on a set of explanatory variables. [4]

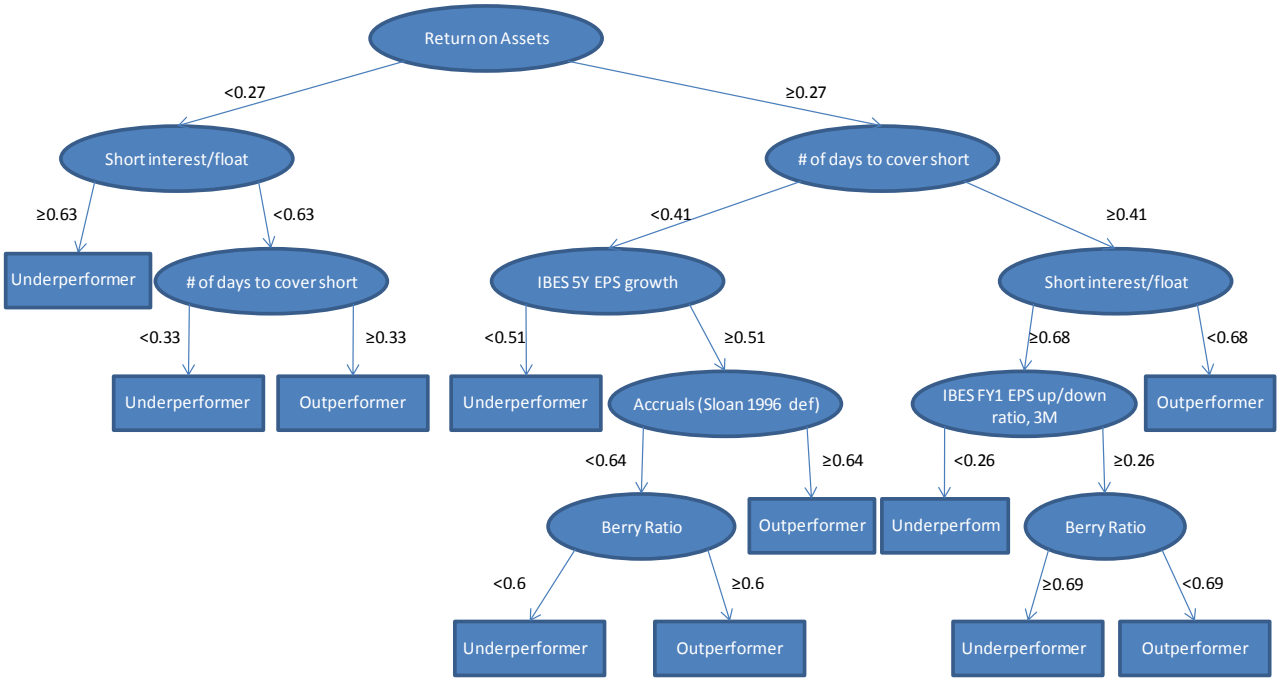
We split our universe into outperformers and underperformers as a classification problem

CART model can easily fit into our stock selection framework problem. We split our universe into outperformers and underperformers as a classification problem. Figure 3 shows a diagram of how a typical CART model works. Essentially, it shows that companies with low return on assets (<0.27) and high short interest (>=0.63) are likely to underperform[5]. In our flagship US stock selection model – QCD model, we developed a innovative way of combining the CART with a more traditional panel data regression model (see Lau et al [2010]).

One issue with decision tree models is that they are prone to overfitting.[6] In practice, we typically overfit the tree model initially, and then prune the redundant branches (via cross validation) to build a smaller but hopefully a more robust model.

[4] We have applied the CART model in many of our past research and found it useful in stock selection (see Luo, *et al* [2010]), news sentiment analysis (see Cahan, *et al* [2010]), quality factor construction (see Cahan, *et al* [2012]), and dividend prediction (see Wang, *et al* [2014a]).
[5] They have a probability of outperformance of less than 50%. Note, in this example, we transform all input factors to a uniform distribution (factors scores are between 0 and 1).
[6] See a detailed discussion on model over-fitting in Wang, *et al* [2012, 2014].

Figure 3: Example of the CART model



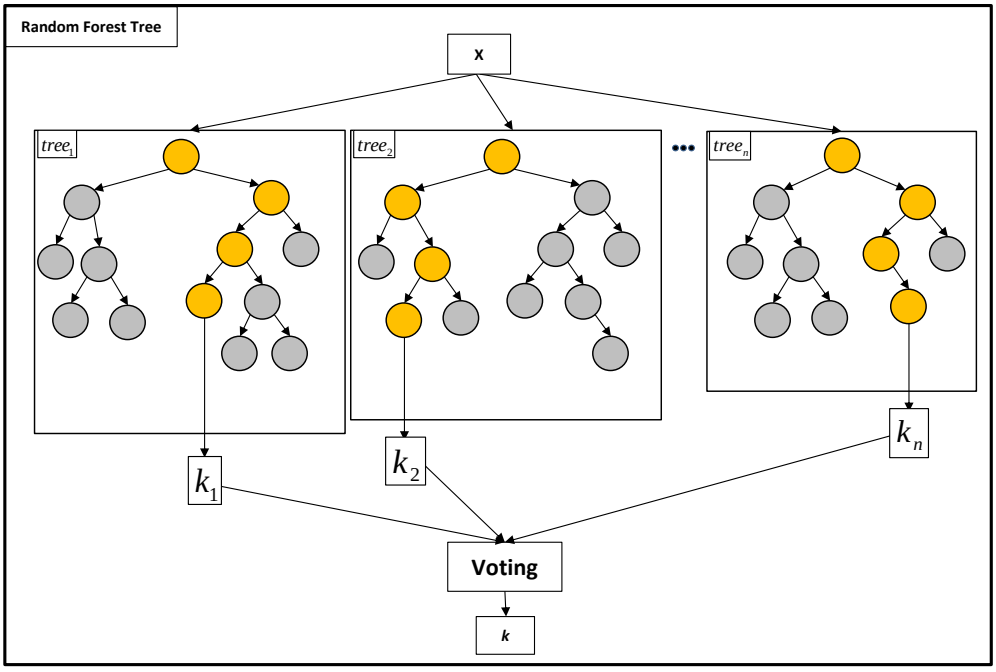
Source: Bloomberg Finance LLP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

Random forest

A natural extension of the CART model is the random forest algorithm. This approach incorporates the idea of "bagging" into the decision tree, i.e., randomly selecting subsets of data and building a tree on each subset. After the training process, the final prediction is the average of the predictions from all the individual trees. The downside of the random forest model (compared to the CART model) is it loses some transparency and is difficult to interpret, especially with hundreds or thousands of trees. An illustration of the random forest tree is shown in Figure 4.

This approach incorporates the idea of "bagging" into the decision tree, i.e., randomly selecting subsets of data and building a tree on each subset.

Figure 4: Random forest

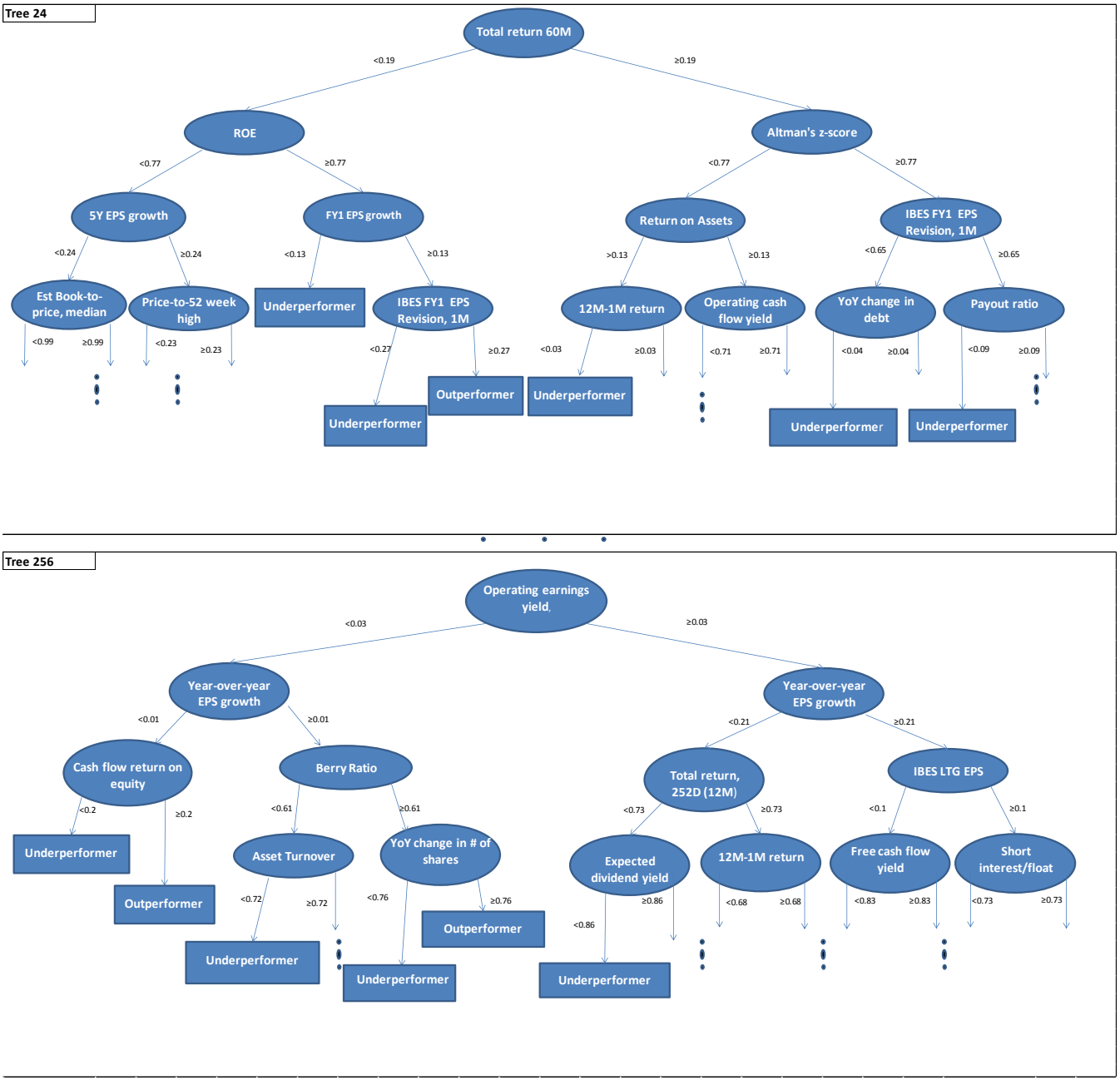


Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

The advantage of the random forest lies in its robustness, compared with the CART models. CART models attempt to capture the most important feature in the data with only a few splits (factors), while in the random forest model; each tree can employ different factors. A typical random forest may have hundreds or thousands of trees, which tends to better capture data patterns, at the cost of interpretability.

Figure 5 shows an example of the random forest model. There are hundreds of trees in this forest. We plot two sample underlying trees. Each tree has many layers. Each tree can be interpreted the same way as our CART model in the previous section.

Figure 5: Example of the random forest



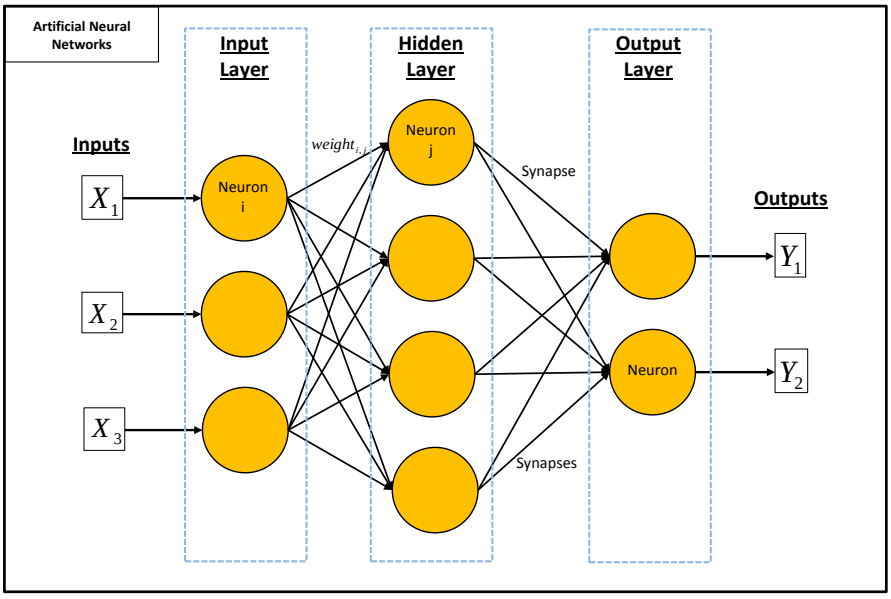
Source: Bloomberg Finance LLP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

Neural Networks

Another popular machine learning technique often used in finance is artificial neural networks (ANN). Essentially you have a series of inputs and outputs and the neural network will tell you what inputs best predict the output. The most commonly used ANNs typically have three layers: the input layer with neurons, which send data via synapses to the hidden layer of neurons, and then via more synapses to the output layer of neurons (see Figure 6 for an example).

Essentially you have a series of inputs and outputs and the neural network will tell you what inputs best predict the output

Figure 6: Artificial neural networks



Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

For the neural networks, the input $\boldsymbol{X}_i$ is our factors, and outputs $\boldsymbol{Y}_i$ is the label of outperformers and underperformers. An ANN model can have more than three layers. As more layers are introduced, in-sample performance typically increases, but out-of-sample performance may suffer, i.e., overfitting.

There are two significant downsides to the ANN model. First, it can be easily over-fitted to show great in-sample performance, but out-of-sample performance tends to be poor. Second, ANN models are also notoriously known for their lack of interpretability, i.e., what do the hidden layers really mean?

Deep learning

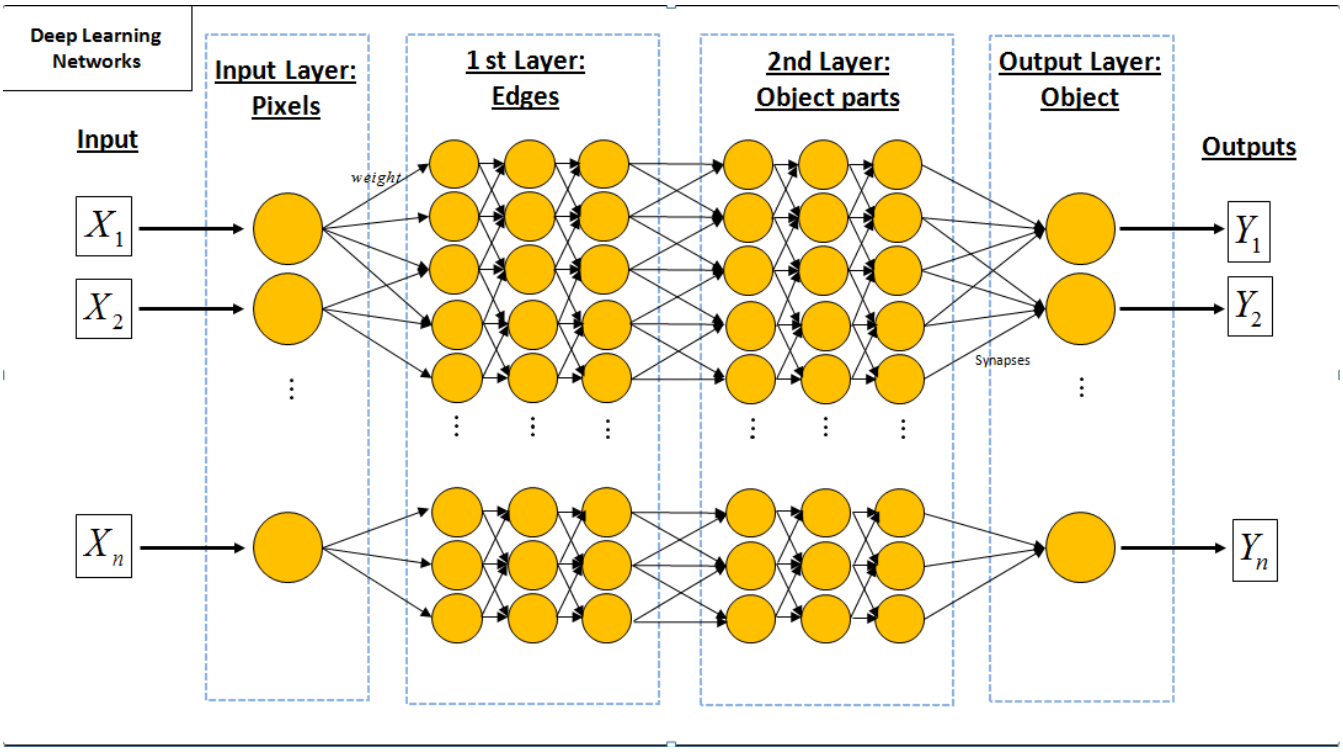
Deep learning[7] is a branch of machine learning that attempts to model the abstractions in data by using hierarchical layers of complex structures. Deep learning originates from artificial neural networks with multiple layers. Tremendous increasing in computing power and availability of data in recent years made it possible to train the deep learning networks with robust performance.

Deep learning originates from artificial neural networks with multiple layers

The algorithm works similar to the human perception process. Humans tend to learn using multiple layers of representation corresponding to different levels of abstraction. The higher level more abstract concept is being learned from lower level raw data. As shown in the Figure 7, we illustrate how deep learning recognizes objects.

[7] Note that the algorithm used by Google that beat a professional Go player recently was based on Deep learning

Figure 7: Deep learning



Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

The input is the lower level pixel in the image. The first layer learns the abstract concept of the 'edges' from the pixels. The second layer further abstracts the concept of the 'object parts', and finally the output layer recognizes the 'objects' from the 'object parts'. For each layer, there can be multiple layers of neutral networks. GPU[8] clustering is usually utilized in order to facilitate the intensive calculation, because of the efficiency of the parallel computing using GPU.

Deep learning has been applied very successfully in the fields of computer vision, speech recognition, natural language processing, and bioinformatics. In the finance world, there are very few applications on deep learning thus far. Deep learning is somewhat a black box and the technology is relatively new. Additionally, asset return prediction tends to be much noisier than other applications of deep learning. However, quants are becoming more interested in utilizing deep learning for stock selection. Some potential applications include:

- Generating sentiment data. Deep learning has proven to be very effective in natural language processing applications.
- Predicting price movement. There has been some work on adapting deep learning methods for time series predictions.

Deep learning has been applied very successfully in the fields of computer vision, speech recognition, natural language processing, and bioinformatics. In the finance world, there are very few applications on deep learning thus far

[8] GPU or graphic processing unit is a specialized electronic circuit designed to rapidly manipulate and alter memory to facilitate image display. Modern GPUs can be more powerful than traditional CPUs for data analysis.

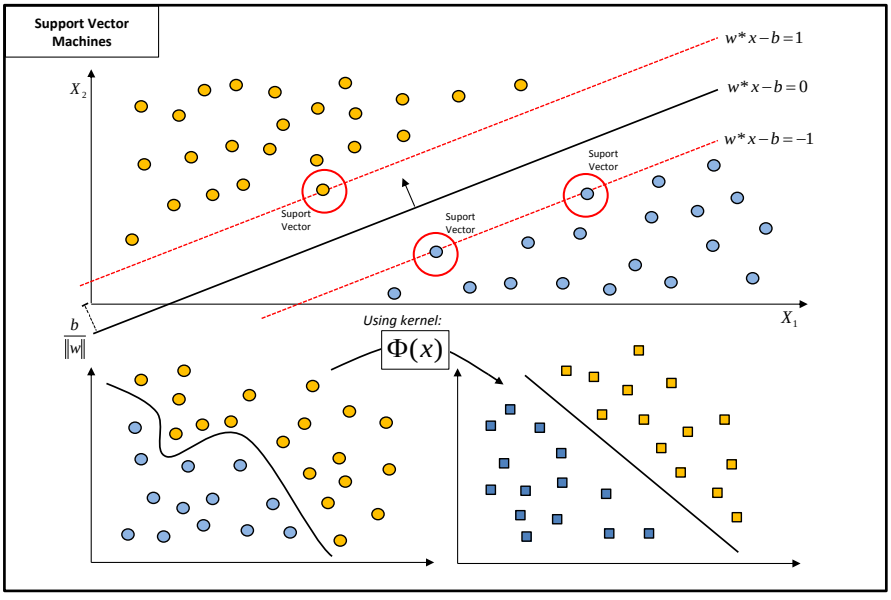
- Analyzing new datasets". As large, non-traditional data sets emerge, deep learning could be a natural tool for processing such complex datasets.

Support vector machine (SVM)

Support vector machine or SVM has gained tremendous reception in technology and finance in recent years[9]. The idea underlying SVM is fairly straightforward. For example, let's say we are trying to predict which stocks outperform or underperform based on two factors or fundamental data items. We plot the factors on a two dimensional plane, giving one color to outperformers and a different color to underperformers (see Figure 8). The idea of SVM is to find a line that can best separate the data points based on their color (i.e. outperformance or underperformance)

The idea of SVM is to find a line that can best separate the data points based on their color (i.e. outperformance or underperformance)

Figure 8: Support vector machine



Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

Any new input data will be then be plotted onto the two dimensional plane. The position on the two dimensional plane will be analyzed to determine if its above or below the computed line which will in turn determine if it is likely to outperform or underperform. It's worthwhile to note that points far away from the line are more likely to outperform or underperform than points closer to the line whose classification is less certain.

The line or function separating the underperformers from the outperformers may not necessarily be a linear function depending upon the data. It could be a nonlinear function.[10] Additionally, in real applications, a vector that can split the two classes perfectly almost never exist. Therefore, a soft margin method is often introduced. A soft margin allows miss-classification and chooses the vector as cleanly as possible. A penalty, which measures the degree of misclassification of the data, can also be introduced.

[9] One of the best website for SVM in finance is www.svms.org/finance with hundreds of reference papers.
[10] Analysts can define a kernel function to map the original data to a higher dimensional space for better separation. The mostly often used kernels are the linear kernel and the radial kernel.

AdaBoost

AdaBoost is a very effective machine learning method for classification[11].The idea behind AdaBoost is that it selects factors or data to try and predict outperformance from underperformance. The unique aspect underlying AdaBoost is that it adaptively chooses a set of factors. These factors are constantly being tweaked to emphasize misclassified stocks. This slowly improves the classification of stocks that would normally be incorrectly classified. Although certain classifiers can be weak, as long as their performance is not random, the performance of the final model will improve. AdaBoost is one of the main algorithm behind our flagship global stock selection model – LASR (see wang, et al [2012, 2013, 2014])

The unique aspect underlying AdaBoost is that it adaptively chooses a set of factors

Regression Models

A logistic regression is a simple classification model. As such it nicely fits in the framework of supervised learning. Essentially, you start with a pool of independent variables that attempt to explain a binary outcome. For example, in the world of finance, you can utilize fundamental data to predict whether a stock outperforms or underperforms. In the world of social media market, you may utilize user characteristics such as age, gender, preferences to try and predict whether a user will purchase an app or not. A logistic regression can be generalized to accommodate more than two outcomes. Alternatively, a decision tree model can accommodate more than two outcomes. We use the logistic regression model in a suite of our research, e.g., dividend growth prediction (see wang, et al [2014]), accounting fraud detection (see Jussa, et al [2015]), and shareholder activism (see Jussa, et al [2016]).

Which machine learning algorithm performs the best?

One of the questions we often encounter is which machine learning model tends to work the best. Undoubtedly, it depends on the application, the problem you are trying to solve and the underlying data.

To address this question, we compared five of the most popular machine learning algorithms in terms of computational speed, model performance and transparency for stock selection, Figure 9. A green light is in favor of a model, while a red light is a warning signal. [12] Based on our analysis, the AdaBoost model performs the best in terms of transparency, performance, and computational time. Again, we would like to warn the careful readers that they should not generalize the conclusion. The choice of machine learning algorithm is very context specific. Researchers should try and compare multiple algorithms before making a final decision.

Figure 9: Comparison for different machine learning algorithms

Machine learning algorithm	Transparency	Performance	Computing time
CART(classification and regression tree)	🟢	🔴	🟢
Random forest	🔴	🔴	🔴
ANNs (Artificial neural network)	🔴	🔴	🟢
SVM (Support vector machine)	🔴	🟢	🔴
AdaBoost	🟢	🟢	🟢

Source: Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope, Deutsche Bank Quantitative Strategy

[11] See Schapire, *et al* [1998] and Wang, *et al* [2010])
[12] See Wang, *et al* [2014b],

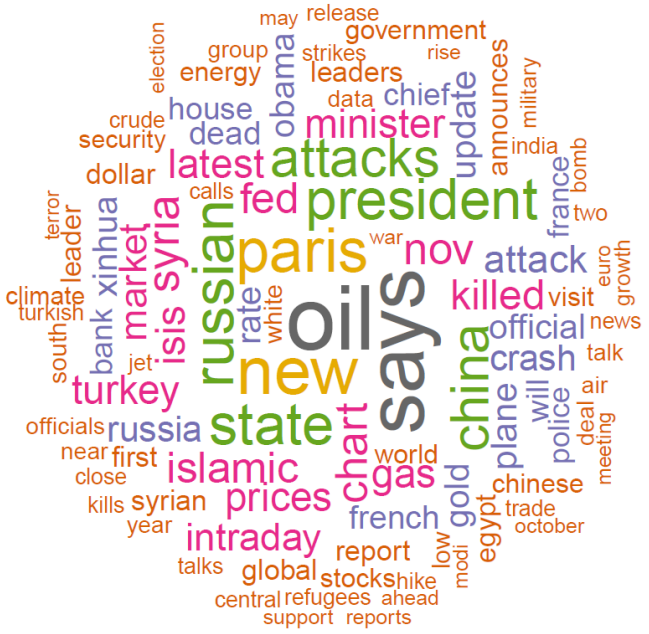
Natural language processing on textual data

The advancement in natural language processing (NLP) techniques had made possible the incorporation of non-numerical or unstructured data in decision making. While market participants are adept in interpreting numbers, digesting millions of news articles requires more time and efforts, both scarce commodities in the investment world. Therefore, a text mining system that extracts signals from textual news can potentially improve the predictive power of a trading model by incorporating this untapped data source.

We use R for text mining. Although the packages for NLP in R are still developing, we found a few of them useful. You can customize packages to your needs and add new frameworks as desired. Figure 10 shows a word cloud for some macroeconomic news events. Size of each word here is related to the frequency of occurrence.

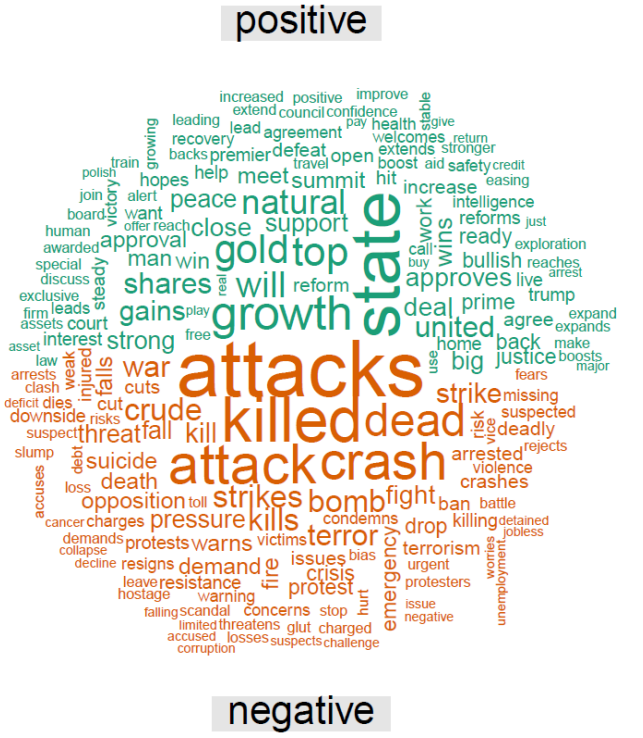
We use R for text mining

Figure 10: Macroeconomic word cloud



Source: RavenPack, Compustat, Capital IQ, Thomson Reuters, Deutsche Bank Quantitative Strategy

Figure 11: Word sentiment cloud



Source: RavenPack, Compustat, Capital IQ, Thomson Reuters, Deutsche Bank Quantitative Strategy

You can also create a sentiment score for news stories, which is a more intricate process, as you need to assign a positive/negative sentiment to each word and then count the occurrence. Creating a sentiment score can be based on seeking out words associated with positive or negative sentiment. Missed, downgrade, and in-the-red are words associated with a negative sentiment. While words such as beat, upgrade, and bullish are words associated with a positive sentiment.

As can be seen form Figure 11, your sentiment scoring is as good as your word list of positive/negative sentiment words. Word matching or pairing algorithms are often used in NLP. One can employ a more rigorous approach based on machine learning to assign these sentiment scores.

Graph theory

Visual representation provides a fast and intuitive way for human brains to digest large amounts of information. In the age of constant data bombardment, visualization tools are arguably more useful than ever. Good data visualization registers comparative analysis, trends and outliers, thus promoting insight and understanding.

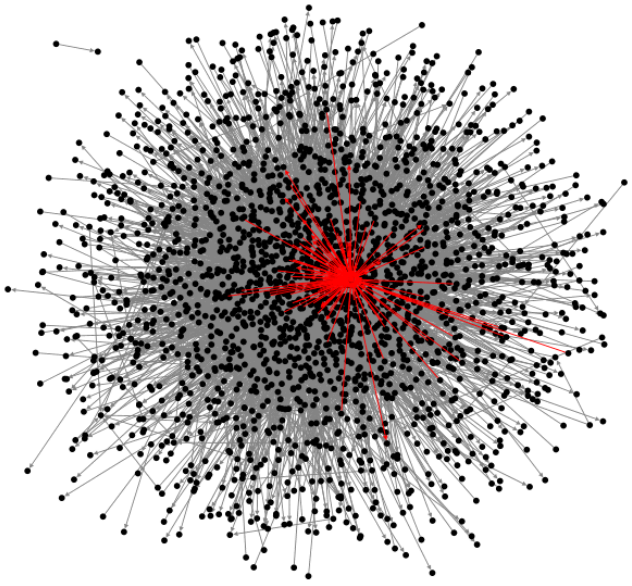
Graph analysis is becoming increasingly important in modeling the multi-dimensional relationships that are common within social media systems, communication networks, electronic circuits, and even biological ecosystems. Graph analytics is the process of interpreting graphical information for the purposes of pattern recognition, data mining, extrapolation, and data linkage.

Graph analysis is becoming increasingly important in modeling the multi-dimensional relationships that are common within social media systems, communication networks, electronic circuits, and even biological ecosystems

A number of data sets can be seen as graphs. A graph is a set of nodes (also called vertices) linked together by edges. These edges can be directed: in this case they represent a flow from one node to another. Undirected edges, on the other hand, represent a form of correlation between two nodes, without a specified direction.

A number of data sets in finance industry naturally fit within the graph theory framework. Supply chain is one example. We have worked on these datasets during our research on supply chain, pairs trading, and asset ownership.[13]

Figure 12: Apple's position within the US supply chain network



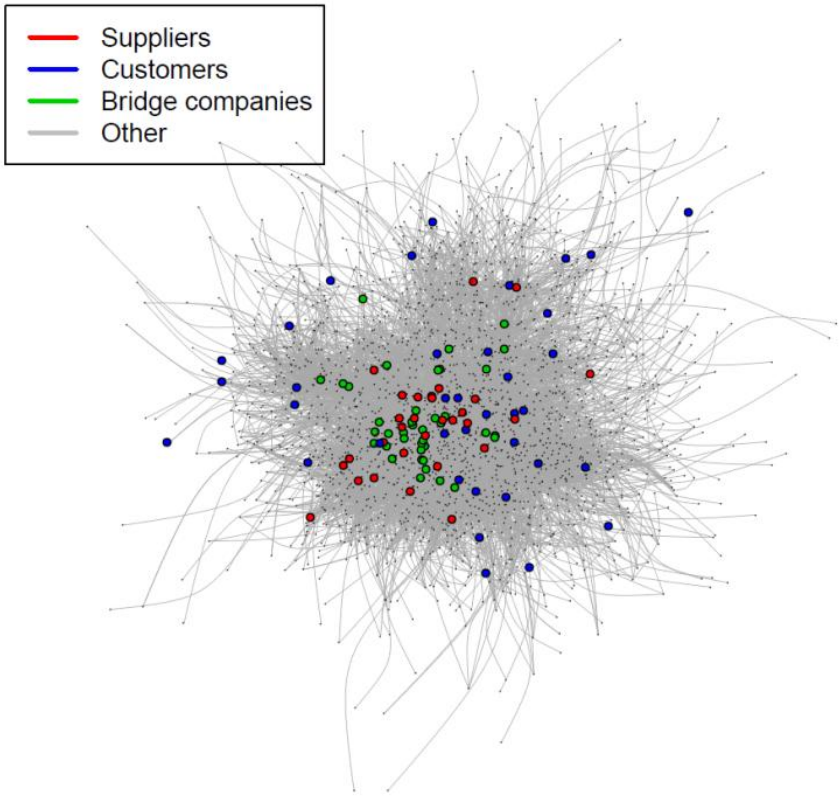
Source: FactSet, Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope,, Deutsche Bank Quantitative Strategy

Beyond data visualization, graph theory brings with it a number of unique and powerful analytical tools. Examples are clustering algorithms and algorithms that detect nodes of interest such as core nodes or chokepoints. A notable model is Google's Pagerank, which detects nodes that accumulate an important amount of inflow. Such models are best at measuring the impact of

[13] See Jussa et al. [2015].

systemic shocks and global trends over the full graph. In our past research, we have used the Pagerank algorithm to detect significant suppliers and significant customers in the supply chain network.

Figure 13: Top suppliers, customers and bridge/chokepoints as of May 2007



Source: FactSet, Bloomberg Finance LP, Compustat, IBES, Russell, S&P, Thomson Reuters, Worldscope,, Deutsche Bank Quantitative Strategy

High frequency

High frequency data does not just stand out for its abundance, but also the analytical tools used to make sense of it. Traditional quant analysis makes heavy use of cross-sectional regressions with a touch of time series analysis. High frequency data, on the other hand, is much heavier on time series models, usually with unique twists to account for idiosyncracies in the data.

The crucial difficulty when applying standard analytical methods onto high frequency data is that the data is *not equally spaced in time*. This is because trades and other events on the order book happen irregularly, usually in bursts of activities with some quiet periods in between. Furthermore, two different stocks may have a very different set of event times and numbers, making direct comparisons difficult. This is called data asynchronicity and has negative effects on correlation estimations and other important analytical tools.

The crucial difficulty when applying standard analytical methods onto high frequency data is that the data is not equally spaced in time

There are usually two approaches to dealing with these irregularity and asynchronicity issues:

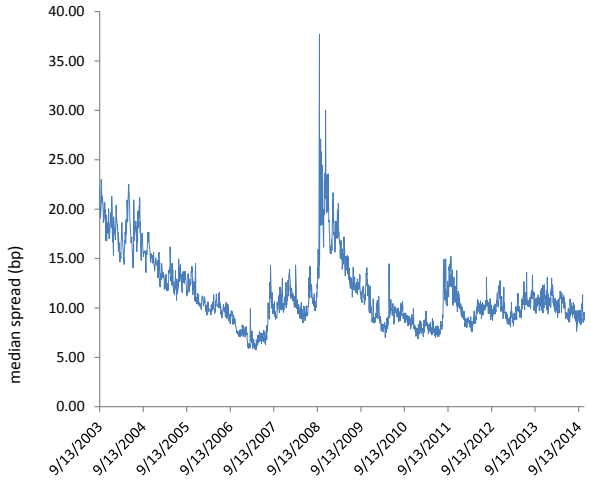
1. measure time in events and not seconds. This is called an event clock.

2. bin the data into intervals: typically one or five minute intervals (i.e., create a periodic time frequency for the data by sub sampling the data every second, for example)

Both methods have their pros and cons and accommodate different types of analysis. Typically, ARIMA, GARCH and other linear time series models work best on the event clock. Volatility and correlation can be more easily estimated with binned data, although specialized methods exist for the event clock[14].

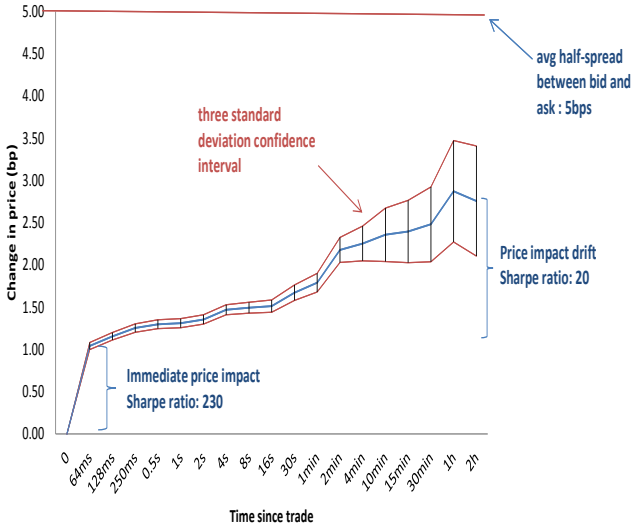
Given the sheer size of high frequency data, a common approach is to use a time series model to summarize the data via a much smaller number of key statistics for the day. These can then be used as factors and compared cross-sectionally at much lower frequencies[15]. The two most important quantities to track in this respect are the bid-ask spread and the price impact of trades.

Figure 14: Daily median bid-ask of Russell 3000 in basis points



Source: TAQ, Deutsche Bank Quantitative Strategy

Figure 15: A visualization of price impact (trade event study) on a specific stock example



Source: TAQ, Deutsche Bank Quantitative Strategy

Multi dimensional

Broadly speaking, quants tend to analyze companies on two dimensions: factors (e.g. consensus analyst estimates for FY1 EPS) and time (e.g. conducting the analysis every month). Several datasets can introduce yet another dimension. For example, analyst detail estimates lay out which analyst made what estimate (FY1 EPS in 2009 made by a XYZ Bank analyst John Doe). Dealing with three dimensional datasets can be computationally intensive.

Dealing with three dimensional datasets can be fairly computer intensive

[14] See High-Frequency Financial Econometrics, Aït-Sahalia & Jacod 2014 for an excellent book on high frequency correlation and volatility estimation
[15] See our previous research in Cahan, et al [2010] and Webster et al. [2015].

The shareholder ownership data is a good example. The 13F ownership data set includes over 50,000 owners, 10,000 stocks over a period of 75 quarters. This represents over 30 billion potential data points. The main challenge presented by such a data set is the tradeoff between two objectives:

1. Easily being able to access, compare and renormalize the data across a given dimension or section of the data. This includes for example re-expressing the data in terms of percentage of market cap owned, percentage of portfolio invested, or comparing shares owned compared to a given segment of stocks or owners.

2. Being able to access the full data set in one session, without having to pre-slice the data or detaching and re-attaching different sections of the data.

For the first objective, the natural data structure is a matrix, or even better, a multidimensional array. For example, one may want to save the data in a 3D matrix format, with the first dimension corresponding to assets, the second to owners and the third to quarters. In this way, renormalizing the data, or collapsing it onto a given segment, is easily achieved by matrix algebra.

However, most of the owners own a tiny amount of the universe in stocks, leading to an enormous amount of sparsity and redundancy in a matrix-representation of the data set. The data is most commonly stored simply as a four dimensional list, with the first column being the stock, the second the owner, the third the quarter and the last one the amount of shares held. Such a data structure typically divides the memory usage by a factor of a hundred. Unfortunately, it also makes manipulating the data highly inefficient for the data scientist.

Our approach to this tradeoff is to use sparse matrices or different size matrices

Our approach to this tradeoff is to use sparse matrices or different size matrices. A package called Matrix in R contains this data structure. These are internally represented in a very compact format similar to the list format described above, while allowing the data scientist to run most matrix operations extremely fast and without thinking much about it. To him or her, the data behaves like a matrix, but all of it fits in memory in a single session. This allows for quick prototyping and backtesting of ideas without compromising the integrity of the data set.

The big aspect of Big data

Historically Big data was out reach for investment managers, given its complexity and unstructured nature. But recently the rapid surge in valuations, buyouts, and IPOs of tech companies has caused a significant increase in entrepreneurial based technology startups hoping to ride this wave. The advancement in computational power and cloud computing environment is also reducing the entry barriers in this space.

Armed with new technologies on NLP, machine learning, graph theory, pattern analysis and machine transcription has helped these start ups efficiently converting Big data to machine-readable forms for systematic purposes. As Big Data world is fast emerging, most of these data vendors are entering into newer space armed with newer technologies.

The advancement in computational power and cloud computing environment is also reducing the entry barriers in this space

Overtime we have integrated and researched several such Big Data vendors .In the ensuing sections we showcase a variety of data structures. We provide a high level summary of the novel aspects of their datasets as well as outline our associated research and analysis. Please contact us at DBEQS.Americas@db.com for more information on any of these datasets.

Satellite Imagery and Pattern Analysis

Objective insights from geospatial data, wireless network signaling data and pattern analysis

Vendors:

- RSMetrics (Satelite Imagery)
- SpaceKnow (Satelite Imagery)
- EidoSearch (Pattern Analysis)
- AirSage (Wireless data)
- Datascription (Text, audio, video AI solution)

Publications

1. *Serving up some clarity on Oil*
2. *DB Quant conference, 2015*
3. *WIP-satellite imaging*

Satellite imagery is taking quant investing to new heights. Companies are providing macro data by processing satellite imagery targeted on ports, roads, ships, pipelines, malls and retail stores. These companies provides data on traffic growth, demographic changes and emerging patterns in goods shipment and purchases.

Another unique source is wireless network signaling data, companies like AirSage generates billions of anonymous location data points in the US, serving market research and other industries for target marketing, equity analytics, site selection, disaster management, consumer research and network analytics.

While EidoSearch provides search and discovery tool built on advanced information processing techniques. It applies pattern matching techniques to security prices and generate forecast of future trends.

Datascription offers an AI solution that discovers, tags and extracts data from textual, audio and video based media. It extracts data using NLP, computer vision, object recognition, machine transcription, amplitude analysis and dynamic corpora compilation. It analyzes output using unique psychographic categories.

Events and Transactions

Unstructured events and financial transaction data such as insider and M&A deals, for predictive analytics

Vendors:

- Capital IQ-Key events and Future events
- Thomson Reuters Deal
- Bloomberg (M&A transaction)
- Capital IQ (Transactions)
- 2iQ (Global Insider Transaction Data)
- Thomson Reuters Activism
- Wall Street Horizon
- Thomson Reuters News Analytics
- RavenPack News Analytics
- Factset Deal Analytics

Wide variety of financial transactions are another source for Big Data analytics. These event based transactions can include insider purchase, M&A deals, private placements, public offerings, buyback and bankruptcies, etc.
Another source of events based unstructured data is companies communication with investors or regualtors. These communications and fillings needs to be structured and scrutinized for specific events, which can be earnings, client or products related. It can also be related to organizational changes such as corporate or executive. Or can be legal or regulatory. In our past research, we extensively covered many of the established vendors in this area these provide high quality data covering most of these event categories. Some vendors like Ravenpack also provide global macro related data. It provides analytical data from Dow Jones, Web and government organizations on economy, currency, regions and commodities.

Publications

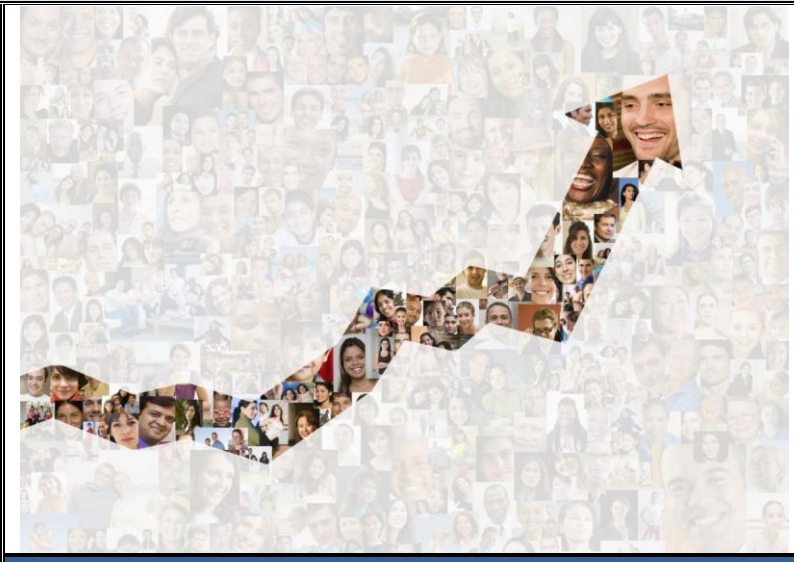
1. *Current Affairs*
2. *Event 2.0*
3. *Event Driven Merger Premia*
4. *Quant 3.0*
5. *Activism, Alpha and Action Heroes*
6. *The Spinoff Premia Wave*
7. *The Curb Appeal of Stock Buybacks*
8. *A Performance Study on Initial PublicOfferings*
9. *Systematic M&A Arbitrage*
10. Beyond Black Litterman
11. Independence Day

Crowd Sourcing

Harnessing the wisdom of the crowd through insights and emerging trends

Vendors:

- Estimize
- Google Trends
- Data Explorer (Markit) - Short Interest



Crowd sourcing is the origin for most big data vendors. The information is collected through websites, surveys, or indirectly as a part of normal business activity of a company.
While some vendors toil hard to source their data, Google has arguable the biggest platform for crowd sourcing. Their Google trends tool based on the Google search engine, provides search-volumes based trend analytics.
Companies like Estimize ask users to submit earnings forecasts for listed companies. The temptation to get noticed encourages the users. Most investors typically rely on providers like IBES, but here is an opportunity to cast the net wider through crowd sourcing.
DataExplorers provide a unique source of securities lending data, it collects information from a wide range of participants in the stock lending market, including beneficial owners, buy side investors, and intermediaries globally.

Publications

1. *Macro Uncertainty, Investor Sentiment, and Asset Returns*
2. *The Long and the Short of it*
3. *The wisdom of crowds*
4. Standing out from the crowd

News Sentiment, Web Mining, and Social Media

Investor mood, sentiment and actionable ideas from web and social media platforms

Vendors:

- RavenPack News Analytics
- Thomson Reuters Newscope
- Alexandria Technology
- Newsquantified
- Market Prophit
- LinkUp
- Accern
- Recorded Future
- Alphasense
- Benzinga

Publications

1. *Beyond the Headlines*
2. *Macro and Micro JobEnomics*
3. *Serving Up Some Clarity on Oil*
4. *Quant 2.0*
5. Quant 3.0
6. Macro Uncertainity, Investor Sentiment, and Asset Returns
7. Surprise!
8. *WIP-Newsquantified*
9. *WIP-RavenPack News Analytics*

World Wide Web with billions of websites and 100's of Exabyte of data is probably the biggest source for crowd sourcing, rich and diversified. Monetizing it is a myriad set of web mining companies.

Web mining companies like Recorded Future, analyze the web for stock trends, cyber threat intelligence, corporate security, competitive intelligence and defense intelligence. Another company in this space is Accern, which provides actionable stories for financial services industry through sentiment analysis, web saturation, source reliability ranks and impact analysis. Market Prophit follows financial bloggers on twitter feeds, systematically ranking financial bloggers and generating sentiment signal at stock level. More specialized companies like LinkUp mines job data for more than 20000 companies from company job websites. Other companies like AlphaSense provides financial search engines relying upon millions of research documents on the web. Their linguistic search and NLP algorithms parse data by topics, concepts and ideas far surpassing basic keyword search.

Benzinga provides actionable news stories and real-time accurate financial data. It also compiles some of the proprietary third party niche datasets mentioned in our report and many more, in a cloud based environment.

Macroeconomic Data

A collection of global macroeconomic indicators and forecasts

Vendors:

- Haver
- Bloomberg Economics
- Datastream
- Bluechip
- Action Economics
- FRED
- World Bank data
- OECD data
- IMF country default, import/export

Publications

1. *Style Rotation*
2. *New Insight in country rotation*
3. *Taper or not, does it matter?*
4. *Quant tactical asset allocation*
5. *Macromomentum Country Rotation*
6. *Country Defaults and Debt Crises*
7. *Independence Day*

At the Macro level there is a wide variety of databases covering different aspects of macro economy. First is a group of government linked public data providers like OECD, FRED, IMF and World Bank. Given their objective is just to relay information to public, these may not provide data in most convenient format and therefore pulling and processing data from these databases can be a painful experience.
Working on this inefficiency are the private organizations like Haver, Action economics, Bluechip and Bloomberg which provide data in a standardized format. This makes it easier to process and integrate in the investment manager's framework.
Data provided can be broad based on economy, currency, trade and countries etc. Or it can also be specialized as the data from IMF on country default or import/export.

Multi-dimensional Data

Supply chain linkage, investor ownership and analyst forecasts

Vendors:

- Thomson reuters Ownership, 13F filing
- Factset Ownership, 13F filing
- Capital IQ Ownership
- Revere - Factset
- Bloomberg supply chain
- Compustat supply chain
- Bureau of Economic Analysis (BEA)
- Capital IQ supply chain
- Thomson Reuters supply chain
- IBES Detail

Databases that are structured and standardized by established vendors can still qualify to be Big Data for their complexity. A traditional example is the IBES detail earnings estimate data which comes with a third dimension of individual analyst recommendation. But since the number of analysts is few, estimates data can be easily analyzed in memory using platforms like R. But for some datasets like Ownership which provides institutional, mutual funds and individual share holder ownership information, the third dimensions can be huge, running in ten's of thousand requiring more sophisticated programs to process.
Yet other databases have complex inter-connected links which makes it tough to process and generate analytical output. Supply-chain is one such database, with numerous links to suppliers, customers, partners and competitors making it way too complex to process and produce insightful outputs.

Publications

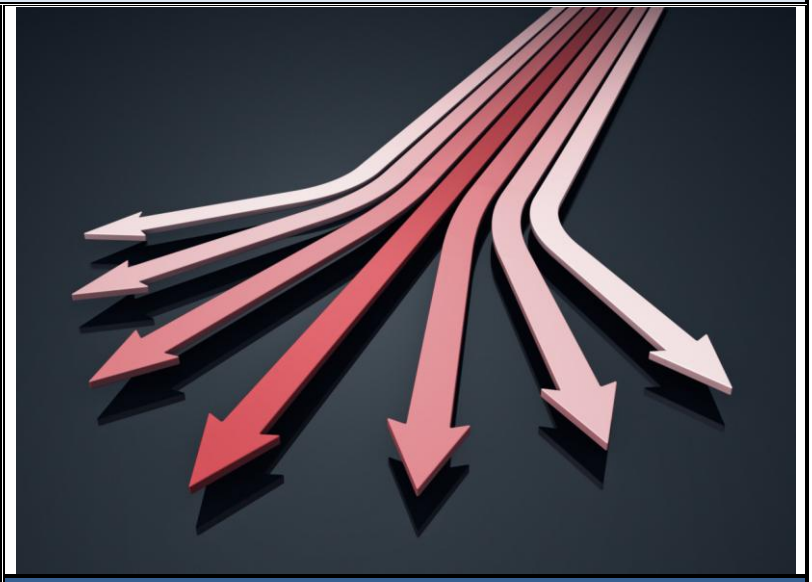
1. *Smart Hedging for Active Management*
2. *Smart Holdings*
3. *Uncovering hidden economic links*
4. *The Logistics of Supply Chain Alpha*
5. *Surprise!*

Cross Asset

Fixed income and options data, global fund flow and hedge fund performance

Vendors:

- Hedge Fund Research (HFR)
- Deutsche Bank eDerivatives database
- OptionMetrics
- Fixed income database (DBIQ)
- EPFR
- Morningstar Fund flow data
- Liper fund flow data



Publications

1. *Hedge Funds: Selecting the Best of the Best*
2. *The Options Issue*
3. *Do Bonds Know Better*
4. *Cross Asset Class Momentum*

Finally there are cross-asset data providers with data on options and fixed income etc. Raw options data is complex with options trading on hundreds of evolving strike prices and multiple expiry dates for a given company. Organization like OptionMetrics are trying to standardized this information coupled with pre-computed volatility, surface and Greeks analytics.
Fixed income database comes with its own set of problems as each issuers have multiple bonds outstanding, with different interest and principal payments cycles and corporate actions.
Other data sources include HFR, which provides detailed data on global hedge funds, their assets under management and returns. EPFR provides global fund flow and asset allocation data. It tracks the traditional and alternatives funds globally and generates country/sector allocation information.

Industry specific

Integrating the fundamental views in quantitative models

Vendors:

- Captal IQ Industry specific
- SNL
- Reuters Industry Data
- Compustat Industry specific
- Compustat Bank & thrift
- Compustat Bank regulatory

Publications

1. *Industry-specific Factors*
2. *DB Quant Handbook of Energy Stock Investing, Part 1*
3. *A Quant Handbook on REIT Investing*

Industry specific data is an important bridge between fundamental and quantitative analysis. Fundamental analysts research a smaller set of companies with depth, while quantitative analysts invest in a larger set of stocks with less company specific knowledge and information. Most quantitative investors rely upon standardized financial data compatible across sectors but using industry specific data they can leverage the fundamental analyst's approach, which traditionally have been out of reach for Quants.

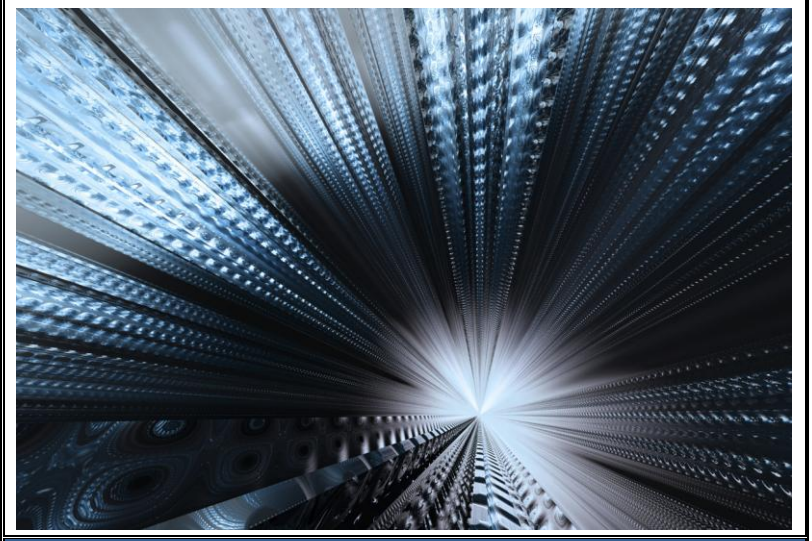
Most established fundamental data vendors provide a set of industry specific factors. But companies like SNL are more specialized in this space, with sector-specific templates and extensive coverage on industry related factors.

High frequency

Tick by tick data for low latency traders and traditional investors

Vendors:

- TAQ - KDB Tick database - NYSE feed
- Reuters data feed
- Bloomberg
- Ablemarkets
- One Tick



Publications

1. *Frequency Arbitrage*
2. *A Portfolio Manager's Guidebook to Trade Execution*

The proliferation of high frequency trading strategies has brought about a paradigm shift within the investment community. Trading strategies that used to be measured in months or days are now delineated in microseconds. This change breeds new challenges from a technological standpoint.

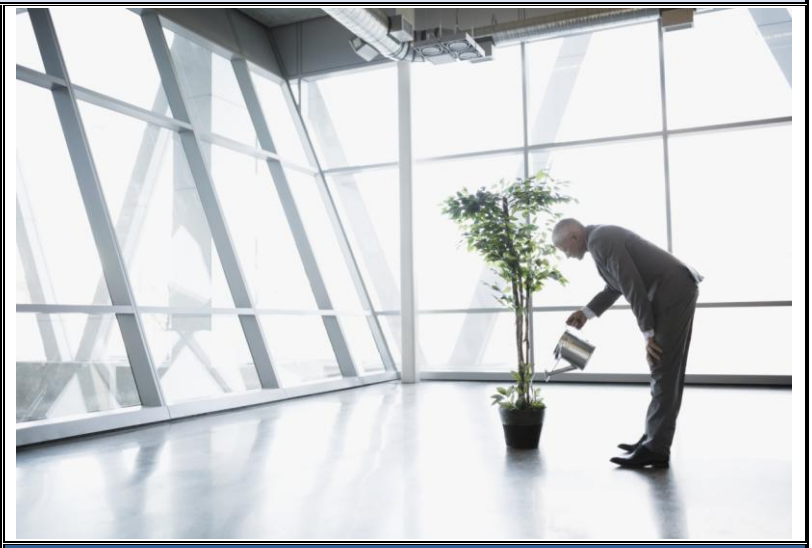
Traditionally this data has been the domain of high frequency traders and agency execution desks but even low frequency investors can extract useful information from high frequency data. Challenge is in the huge volume of high frequency data, that means most participants can't handle it in house. Capitalizing on this inefficiency AbleMarkets provide predictive analytics using high frequency data for portfolio managers, execution traders, risk managers and regulators.

Accounting and Socially Responsible Investing

Sustainable, responsible and impact investing (SRI) through the environmental, social, governance (ESG) criteria

Vendors:

- MSCI ESG, GMI, & AGR
- AAER
- Factset - MSCI GMI/ESG
- Factset - Trucost Environmental Data
- Thomson Reuters Asset4
- Bloomberg ESG (Bcause)
- BizQualify (Labor force)



Publications

1. *The Socially Responsible Quant*
2. *A Darwinian Approach To Detecting Accounting Irregularities*
3. *SRI Integration using Smart Beta*
4. *Forensic Accounting in Global Stock Selection*
5. Accounting for Eighty Million Pensions

As the trend to be more socially responsible is emerging among consumers, entrepreneurs, companies, leaders, and investors alike. Companies are being pushed to integrate socially minded values and principles in their culture, operations, and business practices. As such there has been exponential growth in the popularity for Socially Responsible Investing (SRI).
Established data vendors like MSCI-ESG/AGR are a leading provider of data analytics around this topic and accounting related risks for public companies. MSCI's Accounting and Governance Risk rating (AGR®), is a proprietary rating designed to predict future adverse events such as securities class action litigation, financial restatements, and SEC enforcement actions. AGR encompasses metrics from categories like corporate governance, high risk events, revenue/ expense recognition, and asset-liability valuation. More specialized companies like BizQualify leverages Form 5500 filings with the IRS & Labor Dept., which cover employee benefit plans. This data allows one to track revenue, employee counts, & profitability for private companies using a reliable and verifiable data source.

Big data infrastructure

Concurrent versus sequential architecture

The rapid onset of Big data forces investment managers and data scientists to create as well as adapt to a new analytical toolbox that can grasp, view, and analyze large sets of complex data. In the final section of our Big data primer, we explore and elaborate on the various tools, infrastructure, and programming languages adept to handle potentially unstructured and highly complex datasets.

Since the volume and frequency of data that quants and investors alike have to deal with is rapidly increasing, the concept of concurrent programming is critical. Concurrent computing is an architectural framework where programs are designed to execute in parallel or concurrently. Typically, concurrent programs are designed to run across multiple processors on a single computer or sever. More advanced languages allow support for programs to run across a set of processors or even servers within a distributed network.

Concurrent computing is an architectural framework where programs are designed to execute in parallel or concurrently

One of the main challenges faced by concurrent computing platforms is ensuring the communication and data integrity among various threads or processes. For example, when single process is running in parallel on multiple servers, the architecture must ensure that when data is being passed across the servers, that the data integrity is maintained. Additionally, concurrent computing platforms must coordinate and allocate resources among the various threads. Certain languages employ a shared memory space in order to communicate between multiple threads or sub processes. This method is employed when processes run on the same server or workstation. For distributed concurrent systems, programming languages typically utilize message based communication between threads using a standard network protocol such as TCP/IP.

Programming languages that support concurrent computing have several advantages over traditional programming languages including: increased processing throughput, significant performance boost, more efficient resource utilization, and simultaneous execution of processes.

Specialized computational platforms

The proliferation of Big data sources also requires an evolving set of analysis tools. Below we discuss various Big data analytical tools. By no means is the list of tools fully comprehensive.

Stream based computing platforms

- Stream based computing platforms refer to continuous computations as data flows through a system. Time limitations on data processing are not as stringent as with real-time computational platforms. Distributed stream computing platforms specialize in handling event-driven, real-time data sets that stream data at various, infrequent, and often unbounded rates. These applications typically have the ability to absorb, organize, and statistically analyze large, unconventional, low latency data sets.

- Currently, most stream computing platforms are utilized by trading oriented applications. As these tools continue to evolve, stream based computing platforms will likely be utilized as yet another data processing tool for investors.
- Yahoo's S4 and IBM's System S are both distributed stream computing platforms. S4 is an open source, scalable, plug based platform where users can build algorithms to process continuous data streams. System S is a parallel based stream processing platform that has the ability to filter and correlate large amounts of streaming data from a wide array of sources.

Time limitations on data processing are not as stringent as with real-time computational platforms

Real-time computation platform

- Real time computational systems tend to have extremely strict deadlines where data must be processed. Storm is an open-source distributed real-time computation platform. It provides a simple and easy to use API that works with many popular programming languages. Storm also integrates widely used queuing and database technologies.
- Apache Storm topologies are inherently parallel and run across a cluster of machines. Different parts of the topology can be scaled individually by tweaking parallelism parameters. Apache Storm can process very high throughputs of messages with low latency, and can be used in real-time analytics, online machine learning, and continuous computation.

Real time computational systems tend to have extremely strict deadlines where data must be processed

Advanced data analytical platforms

- The IPython Notebook is a comprehensive and interactive web-based computational environmental. This open source platform supports a host of programming languages including R, Perl, Ruby, Bash, and of course Python. Additionally, the Notebook has support for parallel computing including multicore CPU usage and cloud computing. This toolbox allows for the easy export of results to various formats including LaTex, PDF, and HTML. Version control and live interactive collaboration is also supported. The features and capabilities of web-based computing applications such as the IPython Notebook may be an ideal tool for quant based research, analysis, and collaboration.
- FP Complete, a specialist is an advance computing and processing platform. It enables analysts to perform model experimentation, build prototypes that turn to production codes, and integrate them to existing production system in one seamless development environment. Integrated Analysis Platform (IAP) which is a part of FP Complete offers connectivity to a variety of data formats and services. It can employ the latest Big Data technologies on high performance computing, multicore and distributed functional programming, distributed team management, and cloud management.
- QuantConnect is a startup that offers a cloud-based platform that is pre-loaded with the data and backtesting capabilities. Interestingly, once you are satisfied with your backtested strategy, you can trade it live. It has features to connect your strategy directly to your brokerage account and trade it in real time.

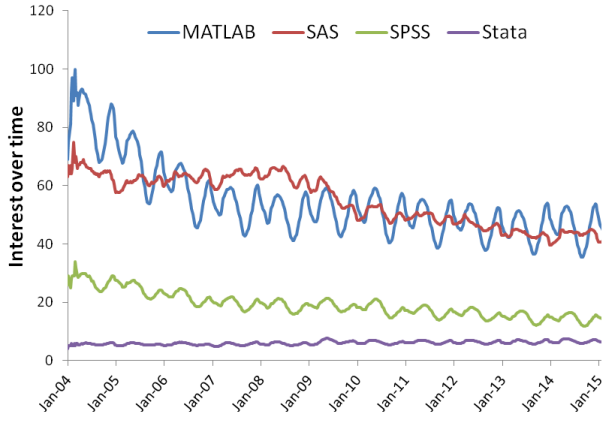
The suite of programming languages

The choice of programming languages is critical when analyzing and working with BigData. You have many choices when selecting a programming language, from lower level languages (Java, C, C++), to middle-level (Python), to higher level (MATLAB, SAS, R). Functionality, speed, memory management, ease of use, development environment, and maintenance must all be considered before selecting a programming language. Most systems support multiple programming languages. Typically database processing is done using a lower level language like SQL or PL/SQL while mathematical and statistical computations are done using higher level languages like R or MATLAB.

Managers must consider the tradeoffs and costs between working with a more mature programming language with strong support and development tools and a more dynamic and cutting edge open source programming language

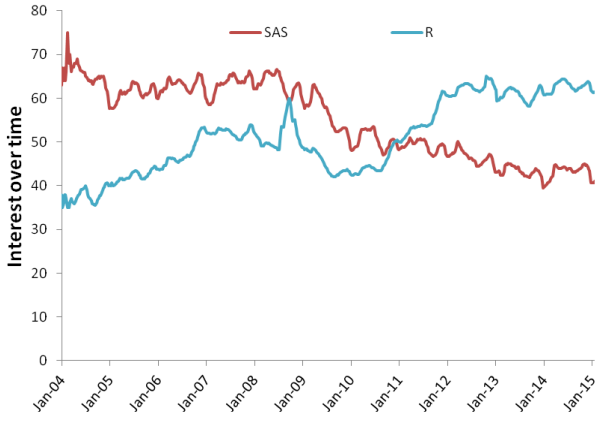
Managers must consider the tradeoffs and costs between working with a more mature programming language with strong support and development tools and a more dynamic and cutting edge open source programming language. Figure 16 and Figure 17 below show the interest level in various programming languages based on Google trends (i.e., Google search volumes). Interestingly, we find an increased growth in more open source and less traditional programming languages. Next, we briefly review the various programming languages and databases, highlighting their inherent advantages as well as disadvantages.

Figure 16: Traditional programming languages trends



Source: Google, Deutsche Bank Quantitative Strategy

Figure 17: R programming language trends



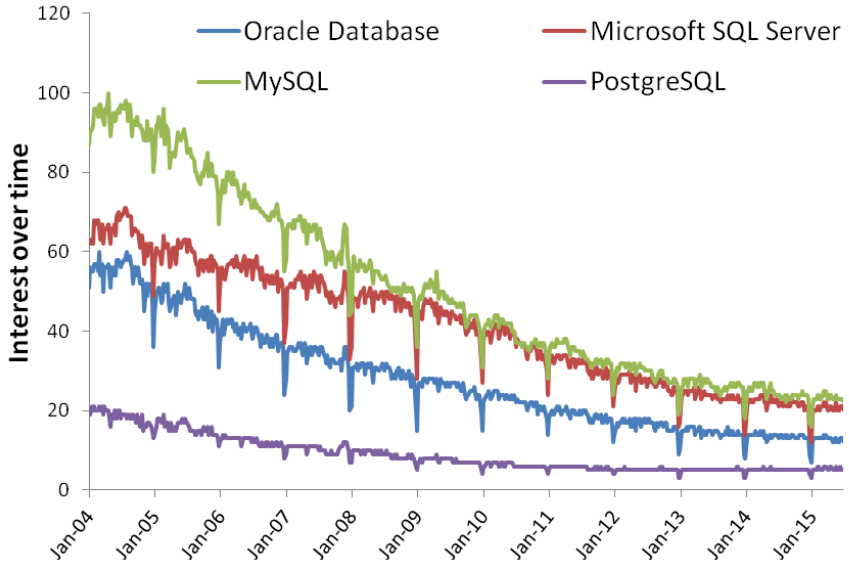
Source: Google, Deutsche Bank Quantitative Strategy

Databases

Relational database management systems (RDBMS)

- For mid to low frequency traditional datasets (i.e., monthly or daily) we prefer a relational database. Two of the most popular relational databases, Oracle & Microsoft SQL Server are both industrial strength and provide good technical support. However they can be expensive and as a result open source databases especially MySQL has gained popularity. But as shown in Figure 18, owning to the Big Data renaissance, RDBMS systems are losing market share at brisk pace.

Figure 18: RDBMS interest over time



Source: Google, Deutsche Bank Quantitative Strategy

- RDBMS systems tend to be optimized for space efficiency, storage and retrieval (i.e., normalized databases). However, most Big data applications require performance rather than efficient storage. As such, flat or de-normalized databases are now becoming more popular in order to handle Big data applications.

Big data databases (scalable distributed file systems)

In the Big data world, databases are essentially a collection of files that can be distributed over multiple systems

- In the Big data world, databases are essentially a collection of files that can be distributed over multiple systems. A processing engine can query all these files and provide data to a client. This architecture is setup for immensely large datasets that require hyper speed access. Traditional databases used a single large, local file to store data and therefore the file required a predefined structure for optimal space efficiency and less redundancy.
- While distributed file systems have a long history, but the publication by Google on GFS (Google File System) in October 2003 vitalized this area of research. The paper by Ghemawat et al [2003], described the architecture and technologies Google developed to handle its thousands of terabytes of data spread across thousands of disks in thousands of machine, concurrently being accessed by large number of clients.
- An example of a distributed file system (DFS) is Hadoop. Most DFSs consists of two parts: a distributed file management system and a distributed processing engine. It splits the files into large blocks and distributes them across nodes in a cluster. It then processes the data. MapReduce is used to process the data. It is composed of two procedures: first Map that does filtering and sorting and Reduce that does the summary or aggregation operations. Overtime DFSs has incorporated many tools for data streaming, data indexing, transformation and in memory processing.
- Other alternatives include document oriented NoSQL databases like MongoDB. It replaces the RDBMS structure with documents with

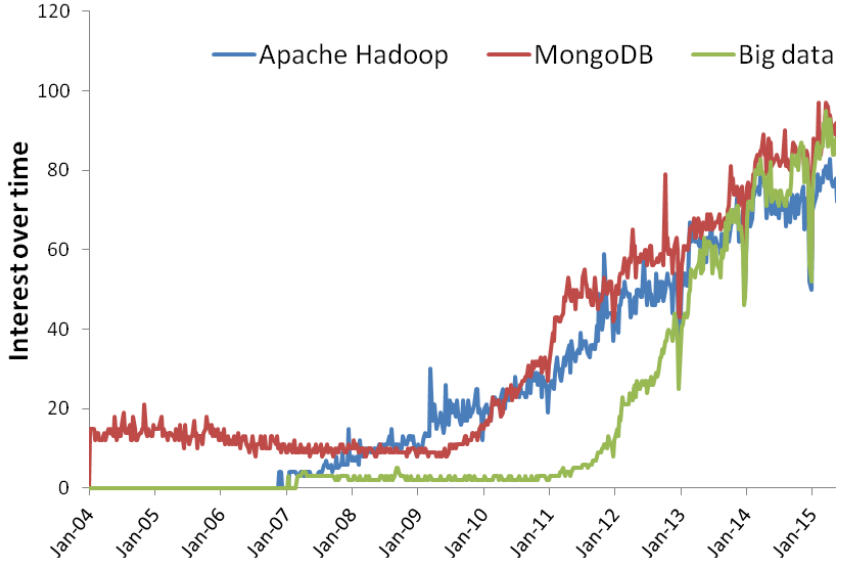
dynamic schemas. These are designed for fast look up through key-value pairs with atomic operations without spanning through records. MongoDB supports Indexing and ad hoc queries such as field and expression searches with many SQL like functionalities.

- NoSQL is a broad category of databases unlike traditional relational databases. Storage and retrieval of data is modeled in means other than the tabular relations used in relational databases. Column, Document, Key-value, Graph and Multi-model based are different varieties of NoSQL databases each with their distinct advantage. The world of NoSQL is fast emerging with a large variety of constantly developing data technologies.

Storage and retrieval of data is modeled in means other than the tabular relations used in relational databases

- In summary, Figure 19 shows the surge in interest in Big Data oriented databases.

Figure 19: Distributed file systems interest over time



Source: Google, Deutsche Bank Quantitative Strategy

High frequency databases

- The volume of high frequency data runs in terabytes and it is almost impossible to manage using a traditional relational database. Fortunately, technology has been keeping pace with the rapidly expanding needs of quants. We use a database called KDB+, it is more suitable for time series data and real-time processing of high frequency data. It has become something of an industry standard for handling tick-by-tick data. KDB+ is one of the new breed of databases specifically designed to hold vast volumes of data in a column-based, in-memory format storing ordered lists sequentially.

The volume of high frequency data runs in terabytes and it is almost impossible to manage using a traditional relational database

- The biggest advantage of this type of databases is speed of access. It comes with its own query language called Q which is able to extract data extremely quickly, and is specifically designed to handle time-series manipulations. The key feature is a proprietary API (built in Java and Q) that dramatically simplifies access to the raw tick data. The API is designed to give researchers a set of tools to do low level data manipulation (e.g., aggregating volume by, say, five minute intervals) without having to write the Q code themselves.

- Another key feature of the API is the ability to call it from R. This means complicated statistical procedures that might be difficult in Q can be coded in R. Other alternatives to KDB are OneTick and Vhayu, these are also optimized for tick data handling with features that allow rapid querying and manipulation of data.

Scripting languages

Traditional programming languages

- SAS and Matlab both include most standard statistical functions, encapsulate an easy to use development environment including GUI, and have strong customer and technical support teams. These languages are typically not open source and thus depending upon the functionality you need, can be costly. They typically have longer development cycles and learning curves. Additionally, it may take time to incorporate new statistical functions and features. However, code readability and error checking tends to be easier.
- In general, these languages are well suited within sciences, engineering, and business management industries. SPSS, Eviews and Stata are also more traditional statistical programming languages.

Open source: R

- Recently the open source languages like R and Python have gained popularity. R is open-source and therefore very cost-effective. One of the biggest reasons we like R is the tremendous amount of user-generated, cutting edge code, available for easy download via "packages".
- R provides extensive functionality in statistics and econometrics through its numerous packages. It is developed by a large community and has a large active user base that regularly contributes new libraries or packages. The latest statistical techniques are typically available in R in couple of months.
- For massive data sets one would more commonly turn to SAS, and for rapid processing one might consider a compiled language like C++. However, advances in R along these fronts are starting to change this. In terms of data volume, new packages that handle data in a smarter way now make it possible to process truly large datasets in. And On the speed front, new parallel processing techniques allow one to significantly speed up execution without additional programming overhead.
- However, there are a few drawbacks of open source languages. There are no standardized libraries and therefore your code may be difficult to read and debug. R does not come with an advanced programming environment. In addition, there is no technical support for R and its development environment is still maturing.

One the biggest reasons we like R is the tremendous amount of user-generated, cutting edge code, available for easy download via "packages"

Python

- In recent few years, Python has grown significantly and it now sports most of the functionalities of R within standardized packages. It also provides the option of calling the R functions from inside Python.
- Learning programming in Python is considered to be simpler but python's GUI is still new and developing. Python is very popular for web development and text mining. Therefore it integrates well as a statistical language for those involved in web development space.

- Python is more popular for web development whereas R is well-suited for advanced statistical and scientific analysis. In order to extract the strengths of both programming languages, one can also use RPy, a very simple but robust interface to R from Python. RPy enables all modules in R to be installed in Python, allowing quant developers to utilize the quantitative aspects of the R programming language within the context of Python. Interfaces such as RPy will likely gain increasing popularity since they can better streamline quantitative systems and applications.

Python is more popular for web development whereas R is well-suited for advanced statistical and scientific analysis

Memory limitations

- Each programming language has different techniques of dealing with memory management. Languages that load a dataset into memory can process the dataset much faster. However, the size of your dataset is limited by memory and therefore you may need to bulk process your dataset and workflow. Other programming languages swap large datasets between disk and memory. This tend to be slower because disk storage is slow. But this will enable the processing of much large datasets.
- R and Matlab load your data set into the system memory. SAS swaps large datasets between system memory and disk. Hence SAS can deal with much larger datasets, but, with constrained performance.
- If you are on Unix platforms then cache can be used to enhance the memory. This is essentially same as swapping data between memory and disk. From our experience, using R or a similar programming language combined with powerful stateless computational GRID computers (UNIX environment) can help address memory management issues when dealing with large datasets.

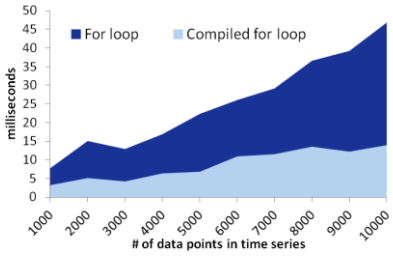
Speeding up heavy computations

Often a program might have a very reasonable memory requirement but the computation might be so heavy that we end up struggling with a slow and cumbersome process. Such cases require a detailed code diagnostic with benchmarking and profiling to highlight the relevant sections with issues. Below are some examples of the techniques which one can employ in R to improve performance.

Compiling to Byte-code

- If your function involves significant arithmetic manipulations, it may be reasonable to compile your code. Compiled languages generally execute faster. Compilers convert your program to machine readable code. Code may run faster but it requires more configuration ands setup depending upon the server you use.
- Interpreted languages like R are executed by an emulator or hypothetical machine. There setup is fairly straight forward. However, R code, can be compiled using the function cmpfun from the library "compiler". You can also use the just in time compiler from library compiler, to compile every function before it runs. Figure 20 plots the performance of an exponential moving average function using compiled and un-compiled for loops.

Figure 20: EMA computation

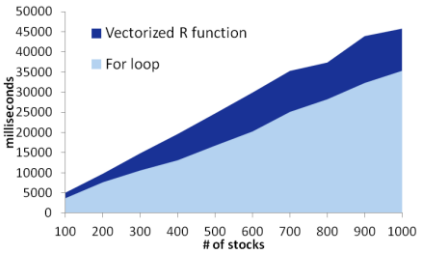


Source: Deutsche Bank Quantitative Strategy

Vectorized functions

- One of the basic concepts in R is vectorization. Variables are created as vectors instead of scalar values in R. A vector is simply an array or of numbers all of same type. This structure allows for R not to perform the low-level operations on each individual scalar value. This makes R code smaller and efficient.
- R provides a family of map functions to which you can apply any arbitrary function to a given set of vectors. Instead of using a traditional "for loop", in R, one can employ a "ply" function. These ply functions are simply for loops optimized to perform an operation on vectors. Figure 21 shows the performance of a computation using for loops and vectorized functions.

Figure 21: z-score computation

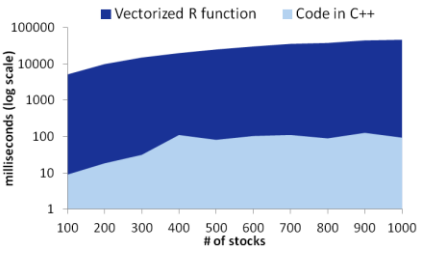


Source: Deutsche Bank Quantitative Strategy

Transforming R code to lower level languages

- In R you have the flexibility to transform the code to C++, a compiled language, by using the cppFunction from library Rcpp. This can have an enormous impact on the performance depending on your program. C++ program can be 100 times faster than R, as in Figure 22.

Figure 22: z-score computation

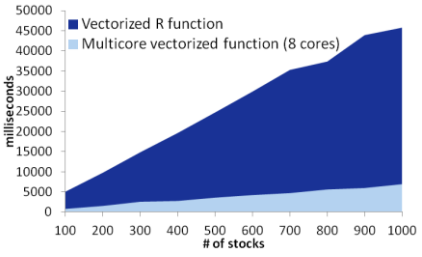


Source: Deutsche Bank Quantitative Strategy

Multi core processing

- Given that most machines are now equipped with multiple processors, it makes sense to utilize them. R is able to spread a complex calculation onto multiple processors for faster computation. R utilizes the following packages for multiprocessing: multicore, snow, doparallel, foreach, and plyr. A simple quant computation was seven times faster using multi core processing in R than without, Figure 23.

Figure 23: z-score computation

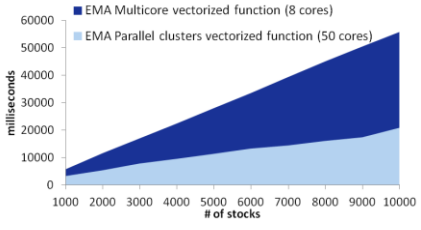


Source: Deutsche Bank Quantitative Strategy

Parallel processing across multiple servers

- The final frontier in process optimization is cloud computing. A stateless grid environment with password-less secure shell (SSH) connection allows us to aggregate cores from multiple servers in a cloud like environment. Figure 24, shows the performance difference for an eight processor (a single server) versus a 50 processor cluster (multiple servers) for an exponential moving average operation.

Figure 24: exp. MA computation using parallel processing



Source: Deutsche Bank Quantitative Strategy

Cloud computing

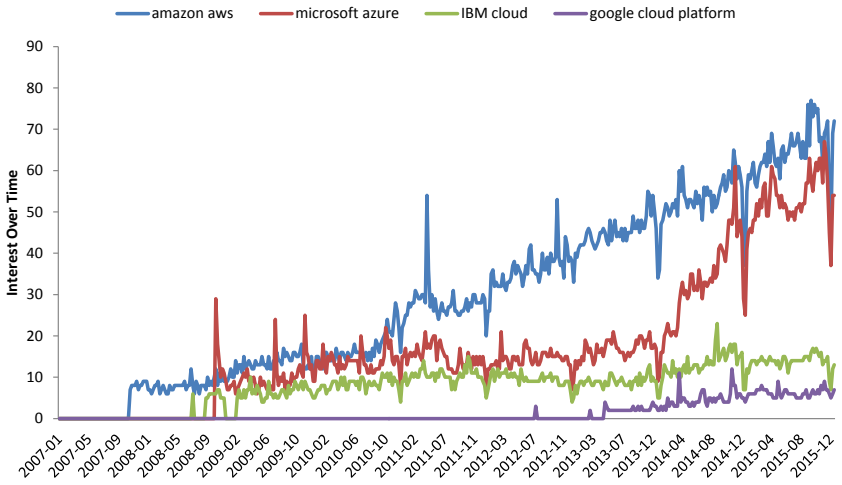
Cloud computing is an increasingly popular technology that involves delivering hosted services over the Internet. A cloud database takes this technology one step further by offering a fully transparent, hosted database solution within the "Internet cloud". Imagine being able to access a fully functional database and supporting infrastructure without having to worry about the complexities of database administration, infrastructure setup, and hosting. Most cloud database solutions offer transparent scalability and administration, automatic performance tuning, backup and disaster recovery, security and user management and authentication.

We think that cloud database systems are particularly useful to quantitative investors because this type of database solutions reduces the development time dramatically. With the proliferation of potential alpha sources, it is critical for quantitative investors to be able to set up and use new datasets in an efficient and timely manner. Cloud databases outsource the complexities of database setup, hardware, and maintenance. In addition, cloud databases reduce overall IT costs by using hardware infrastructure more efficiently.

Cloud database technology of course does not come without its perils

Cloud database technology of course does not come without its perils. Database security and data latency are among a few issues that need to be addressed when hosting data in a cloud. There are several providers of cloud database solutions including SQL Azure (offered by Microsoft), database.com (offered by Salesforce.com), SimpleDB (offered by Amazon Web Services), and App Engine (offered by Google). Figure 25 shows the interest level based on Google trends of various cloud computing platforms.

Figure 25: Interest in Big Data Cloud computing platforms



Source: Google, Deutsche Bank Quantitative Strategy

Many of these providers offer fully transparent cloud database solutions with support for triggers and stored procedures as well as the ability to support a multitude of staging environments including sandbox, test, and production. Figure 26 compares a few cloud computing vendors. Currently, most cloud database solutions are geared towards social and mobile enterprise applications; however, we feel that in the near future, cloud computing and databases could become an advantageous technology

Figure 26: Popular cloud computing platforms

	Compute Options	Storage Types	Object Storage	Firewall/ACL	Pricing Models	Purchase Models
AWS	38	temporary	S3	Yes	Per hour – rounded up	On demand, reserved, spot
GCE	18	temporary/persistent	Google Cloud Storage	Yes	Per minute – rounded up	On demand – sustained use
Azure	33	temporary	Block Blobs	Yes	Per minute – rounded up commitments	On demand – short term commitments

Source: Google, Microsoft, Amazon, Deutsche Bank Quantitative Strategy

Bibliography

Alvarez, M., Jussa, J., Luo, Y., Wang, S. and Wang, A. [2014]. "Hedge Funds: Selecting the Best of the Best", *Deutsche Bank Quantitative Strategy*, Jul 29, 2014

Aït-Sahalia, Y., and Jacod, J. [2014]. "High-Frequency Financial Econometrics", *Princeton University Press*, 2014

Cahan, R., Chen, Z., Luo, Y., Alvarez, M., Jussa, J., and Wang, S. [2012]. "Quant 3.0", *Deutsche Bank Quantitative Strategy*, Dec 17, 2012

Cahan, R., Chen, Z., Wang, S., Luo, Y., Alvarez, M., and Jussa, J. [2013]. "Uncovering hidden economic links", *Deutsche Bank Quantitative Strategy*, Mar 28, 2013

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., Chen, Z., and Wang, S. [2012]. "Standing out from the crowd", *Deutsche Bank Quantitative Strategy*, Jan 31, 2012

Cahan, R., Luo, Y., Jussa, J., and Alvarez, M. [2010]. "Beyond the Headlines", *Deutsche Bank Quantitative Strategy*, Jul 19, 2010

Cahan, R., Luo, Y., Jussa, J., and Alvarez, M. [2010]. "Frequency arbitrage", *Deutsche Bank Quantitative Strategy*, Nov 10, 2010

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., and Weiner, S. [2010]. "The options issue", *Deutsche Bank Quantitative Strategy*, May 12, 2010

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., Chen, Z., and Wang, S. [2011]. "Quant 2.0", *Deutsche Bank Quantitative Strategy*, Nov 18, 2011

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., and Chen, Z. [2011]. "The long and the short of it", *Deutsche Bank Quantitative Strategy*, Jan 18, 2011

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., and Chen, Z. [2011]. "Do bonds know better?", *Deutsche Bank Quantitative Strategy*, May 4, 2011

Elledge, D., Luo, Y., Alvarez, M., Jussa, J., Wang, S., Rohal, G., and Wang, A. [2015]. "Forensic Accounting in Global Stock Selection", *Deutsche Bank Quantitative Strategy*, May 14, 2015

Ghemawat, S., Gobioff, H. and Leung, S. [2003]. "The Google File System", *Google*, Oct 19, 2003

Jussa, J., Alvarez, M., Chen, Z., Wang, S., and Luo, Y. [2013]. "A Performance Study on Initial Public Offerings", *Deutsche Bank Quantitative Strategy*, Oct 03, 2013

Jussa, J., Alvarez, M., Rohal, G., Luo, Y., Wang, S., Wang, A and Mercado, S. [2014]. "Event Driven Merger Premia", *Deutsche Bank Quantitative Strategy*, Sep 2, 2014

Jussa, J., Alvarez, M., Wang, S., Luo, Y., Wang, A, Rohal, G., Elledge, D. and Mercado, S. [2014]. "The Spinoff Premia Wave", *Deutsche Bank Quantitative Strategy*, Nov 4, 2014

Jussa, J., Alvarez, M., Wang, S., Wang, A., Luo, Y. and Chen, Z. [2014]. "Smart Holdings", *Deutsche Bank Quantitative Strategy*, Feb 14, 2014

Jussa, J., Cahan, R., Alvarez, M., Luo, Y., Chen, Z. and Wang, S. [2012]. "Cross Asset Class Momentum", *Deutsche Bank Quantitative Strategy*, Nov 5, 2012

Jussa, J., Cahan, R., Alvarez, M., Wang, S., Luo, Y., and Chen, Z. [2013]. "The Socially Responsible Quant", *Deutsche Bank Quantitative Strategy*, Apr 24, 2013

Jussa, J., Rohal, G., Luo, Y., Alvarez, M., Wang, S., Wang, A. and Elledge, D. [2015]. "A Darwinian Approach To Detecting Accounting Irregularities", *Deutsche Bank Quantitative Strategy*, Mar 4, 2015

Jussa, J., Wang, S., Rohal, G., Alvarez, M., Luo, Y., Wang, A, Elledge, D. Richardson, S., Todd, R. and Aggarwal, S. [2015]. "Serving Up Some Clarity on Oil", *Deutsche Bank Quantitative Strategy*, Feb 3, 2015

Jussa, J., Webster, K., Zhao, G., Luo, Y., Wang, S., Rohal, G., Alvarez, M., Wang, A., and Elledge, D. [2016]. "Activism, Alpha and Action Heroes", *Deutsche Bank Quantitative Strategy*, Jan 6, 2016

Jussa, J., Zhao, G., Webster, K., Luo, Y., Wang, S., Rohal, G., Elledge, D., Alvarez, M., and Wang, A. [2015]. "The Logistics of Supply Chain Alpha", *Deutsche Bank Quantitative Strategy*, Oct 28, 2015

Jussa, J., Zhao, G., Luo, Y., Alvarez, M., Wang, S., Rohal, G., Wang, A., Elledge, D. and Webster, K. [2015]. "Macro and Micro JobEnomics", *Deutsche Bank Quantitative Strategy*, May 19, 2015

Jussa, J., Alvarez, M., Wang, S., Luo, Y. and Chen, Z. [2013]. "SRI Integration using Smart Beta", *Deutsche Bank Quantitative Strategy*, Aug 20, 2013

Luo, Y., Cahan, R., Jussa, J., and Alvarez, M. [2010]. "Style rotation", *Deutsche Bank Quantitative Strategy*, Sep 7, 2010

Luo, Y., Cahan, R., Alvarez, M., Jussa, J., and Chen, Z. [2011]. "Quant Tactical Asset Allocation (QTAA)", *Deutsche Bank Quantitative Strategy*, Sep 19, 2011

Luo, Y., Cahan, R., Jussa, J. and Alvarez, M. [2010]. "Industry-specific factors", *Deutsche Bank Quantitative Strategy*, Jun 7, 2010

Luo, Y., Cahan, R., Alvarez, M., Jussa, J., and Chen, Z. [2011]. "A Quant Handbook on REIT Investing", *Deutsche Bank Quantitative Strategy*, May 2, 2011

Luo, Y., Chen, Z., Wang, S., Alvarez, M., Jussa, J., and Wang, A. [2014]. "Surprise!", *Deutsche Bank Quantitative Strategy*, Mar 10, 2014

Luo, Y., Rohal, G., Alvarez, M., Jussa, J., Wang, S., Wang, A. and Elledge, D. [2015]. "Event 2.0", *Deutsche Bank Quantitative Strategy*, May 12, 2015

Luo, Y., Rohal, G., Alvarez, M., Jussa, J., Wang, S., Wang, A and Elledge, D. [2015]. "Current Affairs", *Deutsche Bank Quantitative Strategy*, Feb 18, 2015

Luo, Y., Wang, S., Cahan, R., Alvarez, M., Jussa, J., and Chen, Z. [2012]. "New insights in country rotation", *Deutsche Bank Quantitative Strategy*, Feb 9, 2012

Luo, Y., Wang, S., Alvarez, M., Jussa, J., Chen, Z., and Wang, A. [2013]. "Taper or no taper - does it matter?", *Deutsche Bank Quantitative Strategy*, Dec 2, 2013

Luo, Y., Wang, S., Alvarez, M., Jussa, J., Wang, A., and Rohal, G. [2014]. "Macro Uncertainty, Investor Sentiment, and Asset Returns", *Deutsche Bank Quantitative Strategy*, Sep 15, 2014

Luo, Y., Wang, S., Alvarez, M., Jussa, J., Rohal, G., Webster, K., Zhao, G., Wang, A., and Elledge, D. [2015]. "Combining Views - Beyond Black-Litterman", *Deutsche Bank Quantitative Strategy*, Sep 14, 2015

Luo, Y., Wang, S., Alvarez, M., Cahan, R., Jussa, J., and Chen, Z. [2013]. "Independence Day", Deutsche *Bank Quantitative Strategy*, Feb 6, 2013

Mesomeris, S., Salvini, M., and Kassam, A. [2010]. "Macromomentum Country Rotation", *Deutsche Bank Quantitative Strategy*, Aug 15, 2010

Rohal, G., Jussa, J., Luo, Y., Alvarez, M., Wang, S., Wang, A and Elledge, D. [2014]. "The Curb Appeal of Stock Buybacks", *Deutsche Bank Quantitative Strategy*, Dec 3, 2014

Wang, A., Alvarez, M., Luo, Y., Jussa, J., Wang, S., Rohal, G. and Elledge, D. [2014]. "Smart Hedging for Active Management", *Deutsche Bank Quantitative Strategy*, Dec 10, 2014

Wang, S., Alvarez, M., Jussa, J., Chen, Z., Wang, A. and Luo, Y. [2014]. "The wisdom of crowds: crowdsourcing earnings estimates", Deutsche *Bank Quantitative Strategy*, Mar 4, 2014

Wang, S., Luo, Y., Alvarez, M., Jussa, J., Wang, A., Rohal, G. and Elledge, D. [2015]. "DB Quant Handbook of Energy Stock Investing", *Deutsche Bank Quantitative Strategy*, Apr 28, 2015

Wang, S., Webster, K., Luo, Y., Alvarez, M., Jussa, J., Rohal, G., Wang, A., Elledge, D. and Zhao, G., [2015]. "Systematic M&A Arbitrage", *Deutsche Bank Quantitative Strategy*, Sep 28, 2015

Webster, K., Luo, Y., Alvarez, M., Jussa, J., Wang, S., Rohal, G., Wang, A., Elledge, D. and Zhao, G. [2015]. "A Portfolio Manager's Guidebook to Trade Execution", *Deutsche Bank Quantitative Strategy*, Jul 8, 2015

Zhao, G., Jussa, J., Luo, Y., Wang, S., Wang, A, Elledge, D. Alvarez, M., Rohal, G., and Webster, K. [2015]. "Country Defaults and Debt Crises", *Deutsche Bank Quantitative Strategy*, Jun 1, 2015

Appendix 1

Important Disclosures

Additional information available upon request

*Prices are current as of the end of the previous trading session unless otherwise indicated and are sourced from local exchanges via Reuters, Bloomberg and other vendors . Other information is sourced from Deutsche Bank, subject companies, and other sources. For disclosures pertaining to recommendations or estimates made on securities other than the primary subject of this research, please see the most recently published company report or visit our global disclosure look-up page on our website at http://gm.db.com/ger/disclosure/DisclosureDirectory.eqsr

Analyst Certification

The views expressed in this report accurately reflect the personal views of the undersigned lead analyst(s). In addition, the undersigned lead analyst(s) has not and will not receive any compensation for providing a specific recommendation or view in this report. Gaurav Rohal/Javed Jussa/Yin Luo/Sheng Wang/George Zhao/Miguel-A Alvarez/Allen Wang/David Elledge

Special Disclosure

An author of this report is on the advisory board of one of the private companies mentioned. The author also has a small equity interest in the company.

Regulatory Disclosures

1.Important Additional Conflict Disclosures

Aside from within this report, important conflict disclosures can also be found at https://gm.db.com/equities under the "Disclosures Lookup" and "Legal" tabs. Investors are strongly encouraged to review this information before investing.

2.Short-Term Trade Ideas

Deutsche Bank equity research analysts sometimes have shorter-term trade ideas (known as SOLAR ideas) that are consistent or inconsistent with Deutsche Bank's existing longer term ratings. These trade ideas can be found at the SOLAR link at http://gm.db.com.

Additional Information

The information and opinions in this report were prepared by Deutsche Bank AG or one of its affiliates (collectively "Deutsche Bank"). Though the information herein is believed to be reliable and has been obtained from public sources believed to be reliable, Deutsche Bank makes no representation as to its accuracy or completeness.

If you use the services of Deutsche Bank in connection with a purchase or sale of a security that is discussed in this report, or is included or discussed in another communication (oral or written) from a Deutsche Bank analyst, Deutsche Bank may act as principal for its own account or as agent for another person.

Deutsche Bank may consider this report in deciding to trade as principal. It may also engage in transactions, for its own account or with customers, in a manner inconsistent with the views taken in this research report. Others within Deutsche Bank, including strategists, sales staff and other analysts, may take views that are inconsistent with those taken in this research report. Deutsche Bank issues a variety of research products, including fundamental analysis, equity-linked analysis, quantitative analysis and trade ideas. Recommendations contained in one type of communication may differ from recommendations contained in others, whether as a result of differing time horizons, methodologies or otherwise. Deutsche Bank and/or its affiliates may also be holding debt securities of the issuers it writes on.

Analysts are paid in part based on the profitability of Deutsche Bank AG and its affiliates, which includes investment banking revenues.

Opinions, estimates and projections constitute the current judgment of the author as of the date of this report. They do not necessarily reflect the opinions of Deutsche Bank and are subject to change without notice. Deutsche Bank has no obligation to update, modify or amend this report or to otherwise notify a recipient thereof if any opinion, forecast or estimate contained herein changes or subsequently becomes inaccurate. This report is provided for informational purposes only. It is not an offer or a solicitation of an offer to buy or sell any financial instruments or to participate in any particular trading strategy. Target prices are inherently imprecise and a product of the analyst's judgment. The financial instruments discussed in this report may not be suitable for all investors and investors must make their own informed investment decisions. Prices and availability of financial instruments are subject to change without notice and investment transactions can lead to losses as a result of price fluctuations and other factors. If a financial instrument is denominated in a currency other than an investor's currency, a change in exchange rates may adversely affect the investment. Past performance is not necessarily indicative of future results. Unless otherwise indicated, prices are current as of the end of the previous trading session, and are sourced from local exchanges via Reuters, Bloomberg and other vendors. Data is sourced from Deutsche Bank, subject companies, and in some cases, other parties.

Macroeconomic fluctuations often account for most of the risks associated with exposures to instruments that promise to pay fixed or variable interest rates. For an investor who is long fixed rate instruments (thus receiving these cash flows), increases in interest rates naturally lift the discount factors applied to the expected cash flows and thus cause a loss. The longer the maturity of a certain cash flow and the higher the move in the discount factor, the higher will be the loss. Upside surprises in inflation, fiscal funding needs, and FX depreciation rates are among the most common adverse macroeconomic shocks to receivers. But counterparty exposure, issuer creditworthiness, client segmentation, regulation (including changes in assets holding limits for different types of investors), changes in tax policies, currency convertibility (which may constrain currency conversion, repatriation of profits and/or the liquidation of positions), and settlement issues related to local clearing houses are also important risk factors to be considered. The sensitivity of fixed income instruments to macroeconomic shocks may be mitigated by indexing the contracted cash flows to inflation, to FX depreciation, or to specified interest rates – these are common in emerging markets. It is important to note that the index fixings may -- by construction -- lag or mis-measure the actual move in the underlying variables they are intended to track. The choice of the proper fixing (or metric) is particularly important in swaps markets, where floating coupon rates (i.e., coupons indexed to a typically short-dated interest rate reference index) are exchanged for fixed coupons. It is also important to acknowledge that funding in a currency that differs from the currency in which coupons are denominated carries FX risk. Naturally, options on swaps (swaptions) also bear the risks typical to options in addition to the risks related to rates movements.

Derivative transactions involve numerous risks including, among others, market, counterparty default and illiquidity risk. The appropriateness or otherwise of these products for use by investors is dependent on the investors' own circumstances including their tax position, their regulatory environment and the nature of their other assets and liabilities, and as such, investors should take expert legal and financial advice before entering into any transaction similar to or inspired by the contents of this publication. The risk of loss in futures trading and options, foreign or domestic, can be substantial. As a result of the high degree of leverage obtainable in futures and options trading, losses may be incurred that are greater than the amount of funds initially deposited. Trading in options involves risk and is not suitable for all investors. Prior to buying or selling an option investors must review the "Characteristics and Risks of Standardized Options", at http://www.optionsclearing.com/about/publications/character-risks.jsp. If you are unable to access the website please contact your Deutsche Bank representative for a copy of this important document.

Participants in foreign exchange transactions may incur risks arising from several factors, including the following: (i) exchange rates can be volatile and are subject to large fluctuations; (ii) the value of currencies may be affected by numerous market factors, including world and national economic, political and regulatory events, events in equity and debt markets and changes in interest rates; and (iii) currencies may be subject to devaluation or government imposed exchange controls which could affect the value of the currency. Investors in securities such as ADRs, whose values are affected by the currency of an underlying security, effectively assume currency risk.
Unless governing law provides otherwise, all transactions should be executed through the Deutsche Bank entity in the investor's home jurisdiction.

United States: Approved and/or distributed by Deutsche Bank Securities Incorporated, a member of FINRA, NFA and SIPC. Analysts employed by non-US affiliates may not be associated persons of Deutsche Bank Securities Incorporated and therefore not subject to FINRA regulations concerning communications with subject companies, public appearances and securities held by analysts.

Germany: Approved and/or distributed by Deutsche Bank AG, a joint stock corporation with limited liability incorporated in the Federal Republic of Germany with its principal office in Frankfurt am Main. Deutsche Bank AG is authorized under German Banking Law (competent authority: European Central Bank) and is subject to supervision by the European Central Bank and by BaFin, Germany's Federal Financial Supervisory Authority.

United Kingdom: Approved and/or distributed by Deutsche Bank AG acting through its London Branch at Winchester House, 1 Great Winchester Street, London EC2N 2DB. Deutsche Bank AG in the United Kingdom is authorised by the Prudential Regulation Authority and is subject to limited regulation by the Prudential Regulation Authority and Financial Conduct Authority. Details about the extent of our authorisation and regulation are available on request.

Hong Kong: Distributed by Deutsche Bank AG, Hong Kong Branch.

India: Prepared by Deutsche Equities Private Ltd, which is registered by the Securities and Exchange Board of India (SEBI) as a stock broker. Research Analyst SEBI Registration Number is INH000001741. DEIPL may have received administrative warnings from the SEBI for breaches of Indian regulations.

Japan: Approved and/or distributed by Deutsche Securities Inc.(DSI). Registration number - Registered as a financial instruments dealer by the Head of the Kanto Local Finance Bureau (Kinsho) No. 117. Member of associations: JSDA, Type II Financial Instruments Firms Association and The Financial Futures Association of Japan. Commissions and risks involved in stock transactions - for stock transactions, we charge stock commissions and consumption tax by multiplying the transaction amount by the commission rate agreed with each customer. Stock transactions can lead to losses as a result of share price fluctuations and other factors. Transactions in foreign stocks can lead to additional losses stemming from foreign exchange fluctuations. We may also charge commissions and fees for certain categories of investment advice, products and services. Recommended investment strategies, products and services carry the risk of losses to principal and other losses as a result of changes in market and/or economic trends, and/or fluctuations in market value. Before deciding on the purchase of financial products and/or services, customers should carefully read the relevant disclosures, prospectuses and other documentation. "Moody's", "Standard & Poor's", and "Fitch" mentioned in this report are not registered credit rating agencies in Japan unless Japan or "Nippon" is specifically designated in the name of the entity. Reports on Japanese listed companies not written by analysts of DSI are written by Deutsche Bank Group's analysts with the coverage companies specified by DSI. Some of the foreign securities stated on this report are

not disclosed according to the Financial Instruments and Exchange Law of Japan.

Korea: Distributed by Deutsche Securities Korea Co.

South Africa: Deutsche Bank AG Johannesburg is incorporated in the Federal Republic of Germany (Branch Register Number in South Africa: 1998/003298/10).

Singapore: by Deutsche Bank AG, Singapore Branch or Deutsche Securities Asia Limited, Singapore Branch (One Raffles Quay #18-00 South Tower Singapore 048583, +65 6423 8001), which may be contacted in respect of any matters arising from, or in connection with, this report. Where this report is issued or promulgated in Singapore to a person who is not an accredited investor, expert investor or institutional investor (as defined in the applicable Singapore laws and regulations), they accept legal responsibility to such person for its contents.

Qatar: Deutsche Bank AG in the Qatar Financial Centre (registered no. 00032) is regulated by the Qatar Financial Centre Regulatory Authority. Deutsche Bank AG - QFC Branch may only undertake the financial services activities that fall within the scope of its existing QFCRA license. Principal place of business in the QFC: Qatar Financial Centre, Tower, West Bay, Level 5, PO Box 14928, Doha, Qatar. This information has been distributed by Deutsche Bank AG. Related financial products or services are only available to Business Customers, as defined by the Qatar Financial Centre Regulatory Authority.

Russia: This information, interpretation and opinions submitted herein are not in the context of, and do not constitute, any appraisal or evaluation activity requiring a license in the Russian Federation.

Kingdom of Saudi Arabia: Deutsche Securities Saudi Arabia LLC Company, (registered no. 07073-37) is regulated by the Capital Market Authority. Deutsche Securities Saudi Arabia may only undertake the financial services activities that fall within the scope of its existing CMA license. Principal place of business in Saudi Arabia: King Fahad Road, Al Olaya District, P.O. Box 301809, Faisaliah Tower - 17th Floor, 11372 Riyadh, Saudi Arabia.

United Arab Emirates: Deutsche Bank AG in the Dubai International Financial Centre (registered no. 00045) is regulated by the Dubai Financial Services Authority. Deutsche Bank AG - DIFC Branch may only undertake the financial services activities that fall within the scope of its existing DFSA license. Principal place of business in the DIFC: Dubai International Financial Centre, The Gate Village, Building 5, PO Box 504902, Dubai, U.A.E. This information has been distributed by Deutsche Bank AG. Related financial products or services are only available to Professional Clients, as defined by the Dubai Financial Services Authority.

Australia: Retail clients should obtain a copy of a Product Disclosure Statement (PDS) relating to any financial product referred to in this report and consider the PDS before making any decision about whether to acquire the product. Please refer to Australian specific research disclosures and related information at https://australia.db.com/australia/content/research-information.html

Australia and New Zealand: This research, and any access to it, is intended only for "wholesale clients" within the meaning of the Australian Corporations Act and New Zealand Financial Advisors Act respectively.
Additional information relative to securities, other financial products or issuers discussed in this report is available upon request. This report may not be reproduced, distributed or published by any person for any purpose without Deutsche Bank's prior written consent. Please cite source when quoting.

David Folkerts-Landau
Chief Economist and Global Head of Research

Raj Hindocha
Global Chief Operating Officer
Research

Marcel Cassard
Global Head
FICC Research & Global Macro Economics

Steve Pollard
Global Head
Equity Research

Michael Spencer
Regional Head
Asia Pacific Research

Ralf Hoffmann
Regional Head
Deutsche Bank Research, Germany

Andreas Neubauer
Regional Head
Equity Research, Germany

International Locations

Deutsche Bank AG
Deutsche Bank Place
Level 16
Corner of Hunter & Phillip Streets
Sydney, NSW 2000
Australia
Tel: (61) 2 8258 1234

Deutsche Bank AG
Große Gallusstraße 10-14
60272 Frankfurt am Main
Germany
Tel: (49) 69 910 00

Deutsche Bank AG
Filiale Hongkong
International Commerce Centre,
1 Austin Road West,Kowloon,
Hong Kong
Tel: (852) 2203 8888

Deutsche Securities Inc.
2-11-1 Nagatacho
Sanno Park Tower
Chiyoda-ku, Tokyo 100-6171
Japan
Tel: (81) 3 5156 6770

Deutsche Bank AG London
1 Great Winchester Street
London EC2N 2EQ
United Kingdom
Tel: (44) 20 7545 8000

Deutsche Bank Securities Inc.
60 Wall Street
New York, NY 10005
United States of America
Tel: (1) 212 250 2500