

# Natural Language Processing – Part I: Primer

## Unveiling the Hidden Information in Earnings Calls


Author  
Frank Zhao  
Quantamental Research  
617-530-8107  
fzhao@spglobal.com

Interest in natural language processing (NLP) has grown in earnest since Turing's publication "Computing Machinery and Intelligence"<sup>1</sup> in 1950. In his seminal work, Turing laid out his criterion for intelligence – a computer could be considered intelligent if it can interact with humans without them ever realizing they were dealing with a machine. NLP at its core is the embodiment of that vision where consumers of NLP obtain useful insights from data without ever needing to know whether they are interacting with a machine.

Given the growing interest in NLP among investors, we are publishing this primer to demystify many aspects of NLP and provide three illustrations, with accompanying Python [code](#), of how NLP can be used to quantify the sentiment of earnings calls. In our first example below, sector-level sentiment trends are generated providing insights around inflection points and accelerations. The other two illustrations are: i) stock-level sentiment changes and forward returns ii) language complexity of earnings calls ([Section 4](#)).

### S&P 500 Earnings Call Sentiment Trends

GICS Sectors	Calendar Quarters			
	Q3 2016	Q4 2016	Q1 2017	Q2 2017
Industrials	-2.6%	5.9%	15.5%	15.7%
Information Technology	4.1%	0.3%	7.2%	15.2%
Utilities	-4.6%	7.2%	7.0%	14.1%
Financials	-3.4%	3.7%	14.1%	13.0%
Energy	0.1%	13.6%	25.8%	10.3%
Materials	5.7%	-1.8%	12.2%	1.4%
Consumer Discretionary	-9.9%	-2.2%	4.5%	0.1%
Consumer Staples	3.2%	-1.9%	-8.0%	-5.0%
Health Care	-5.7%	-2.9%	-1.3%	-7.0%
Telecommunication Services	9.1%	-13.6%	-36.9%	-29.4%



Note: Sentiment is defined as the proportion of negative words in an earnings call using [Loughran and McDonald \(2011\)](#)<sup>2</sup>. Sentiment changes are measured quarter-over-quarter from four quarters ago where the values are multiplied by -1 to make them easier to interpret. Sector-level values are rolled up equally from the stock-level. Real estate is rolled into Financials. Source: S&P Global Market Intelligence Quantamental Research, as of 08/08/2017.

The paper is laid out into four sections:

- **What is NLP?** – We demystify common NLP terms ([Section 2](#)) and provide an overview of general steps in NLP ([Section 2](#)).
- **Why is NLP important?** – Forty zettabytes (10<sup>21</sup> bytes) of data are projected to be on the internet by 2020,<sup>3</sup> out of which more than eighty percent of the data are unstructured in nature, requiring NLP to process and understand ([Section 3](#)).
- **How can NLP help me?** – We derive insights from earnings call transcripts via NLP measuring industry-level sentiment trends or language complexity of earnings calls and much more ([Section 4](#)).
- **Where do I start?** – Code for each use case is enclosed, enabling users to replicate the sentiment analysis ([Section 6](#)).

<sup>1</sup> Turing, A.M. "Computing Machinery and Intelligence." *Mind* 49 (1950), 433-460.

<sup>2</sup> Loughran, T., AND B. McDonald. "When is a Liability not a Liability? Textual analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (2011), 35-65.

<sup>3</sup> Mearian, L. (2012, Dec. 11). By 2020, there will be 5,200 GB of data for every person on Earth. Retrieved from <http://www.computerworld.com>.

## 1. Introduction

NLP dates back to as early as the seventeenth century, when philosophers and polymaths such as Gottfried Leibniz and Rene Descartes put forth theoretical proposals to relate words between languages. The first patents for translating machines were filed in the 1930s by Grigor Artsrouni, an Armenian journalist and writer. In 1950, Alan Turing, the coiner of the Turing Test, published his seminal article “Computing Machinery and Intelligence”. In 1957, Noam Chomsky’s “Syntactic Structures” revolutionized linguistics with his ‘universal grammar’, a collection of rules based off of syntactic structures. Through the 1980s, all NLP systems were primarily based on complex sets of hand-written rules and not until the late 1980s did statistical algorithms start to replace the hand-written rules and gain ubiquity. In the following sections, we try to touch on (almost) every aspect of basic NLP such that by the end of the primer we hope our readers not only have a foundational knowledge of the subject, but also able to perform simple NLP using our shared code snippets.

## 2. What is NLP? – Definitions & Steps

In this section, we demystify a number of NLP related terms and provide our readers with our simple definitions that will serve as a foundation for all subsequent sections. Then, we will walk our readers through the major steps within a NLP process.

### 2.1 Definitions

Terms such as big data, alternative data, machine learning, supervised learning and unsupervised learning are being thrown around in the context of NLP. What exactly do they mean? In fact, what does NLP itself really mean?

#### 2.1.1 Definitions: Overarching Terms

**Natural Language Processing (NLP):** leveraging computers and statistics to process and make sense of language in a systematic and sensible way that doesn’t involve human intervention besides coding.

**Artificial Intelligence (AI):** is the process of building systems (that ingest data, use different learning techniques on the data to create intelligence and then output that intelligence) that can do intelligent – i.e., human-like – things. Our take is that true AI is still some time away, but Google AlphaGo<sup>4</sup> and Facebook<sup>5</sup> seem to have made strides recently. This is the ultimate goal.

---

<sup>4</sup> Mozur, P. (2017, May 25). Google’s A.I. Program Rattles Chinese Go Master as It Wins Match. Retrieved from <http://www.nytimes.com>.

<sup>5</sup> LaFrance A. (2017, June 15). An Artificial Intelligence Developed Its Own Non-Human Language. Retrieved from <https://www.theatlantic.com>.

### 2.1.2 Definitions: Categories of Data

**Structured Data:** a data set that has been cleaned of errors, processed to a standardized format and readily stored in a carefully designed SQL database.

**Unstructured / Big / Alternative Data:** these terms describe a subset of newly created data that hasn't been (or fully) explored, cleaned and processed. It is usually heavy with text, but may also include audio and/or visual data. All data that does not qualify as structured is unstructured.

### 2.1.3 Definitions: Categories of Learning Methods

**Machine Learning (ML):** a set of statistical methods to help make sense of a data set to gain a useful insight or to automate a task. A computer or data scientist uses the term "machine learning" whereas a pure statistician may use "applied statistics" or an economist may use "econometrics". There are two major categories to which all algorithms in machine learning belong: **supervised and unsupervised learning**. See details below.

**Deep Learning (DL):** is a subfield of machine learning. It is most commonly implemented using a neural network that takes its design from the human brain – a network of neurons that are connected by synapses. One way to visualize deep learning via a neural network is that the learning is organized as a cylindrical-shaped hierarchy of concepts, where the concepts could be non-linear. The concepts increase in complexity as one moves from the topmost layer to the bottommost one. The main difference between machine learning and deep learning is that the performance of deep learning algorithms increases linearly as the amount of input data increases whereas the performance of machine learning algorithms plateaus.

**Supervised Learning (SL):** A category of **machine learning** or **deep learning** in which the researcher or data scientist feeds an algorithm some training data such that the algorithm has ex-ante 'rules' (e.g., the data points are linearly related, the data points follow the normal distribution or the data points have certain probabilities of taking on certain values). Examples of supervised learning methods are linear regression, Bayesian statistics and decision trees.

**Unsupervised Learning (UL):** A category of **machine learning** or **deep learning** in which the algorithms themselves find 'hidden' patterns by sorting data points into different groups or categories until the algorithms find an optimal divide. It is like an optimization problem (for those readers who have an operational research background) where the algorithm is trying different groupings until it finds a global or a local 'maximum' or 'minimum'. The main difference from supervised learning is that no 'rules' or training data sets are given. Examples of unsupervised learning methods are clustering, and PCA expectation-maximization.

#### 2.1.4 Definitions: Others

**Tokens:** words or entities (e.g., punctuations, numbers, URLs, etc.) present in a text.

**Tokenization:** the process of converting a text (a list of strings) into tokens.

### 2.2 General Steps in NLP

There are generally three major steps in NLP (see details below): text preprocessing, text to features and testing & refinement. Text preprocessing includes noise removal, lexicon normalization and objective standardization. Text to features includes syntactical parsing, entity parsing, statistical features and word embedding. Lastly, testing & refinement step includes the designing, the calibrating and the refining of a model that is the engine which helps users to extract useful information from a content set or perform an automated task.

#### 2.2.1 Text Preprocessing

**Noise removal** is cleaning of textual data by stripping away anything that is not relevant to the task at hand including but not limited to the removal of acronyms, punctuations and numbers. Usually, URLs, hashtags, acronyms and stopwords (e.g., words like *the*) are removed. In fact, anything that isn't relevant to one's analysis could be considered as noise. One solution that we used is to come up with a defined set of words and entities such that everything else was removed as noise.

**Lexicon normalization** is the method of standardizing multiple representations that are exhibited by a single word (e.g., different inflections of a word: e.g., play, plays, played, etc.). Two commonly used solutions are stemming and lemmatization. The goal of both is to reduce inflectional forms and derivationally related forms of a word to its root. Stemming is a cruder process of chopping off the suffix of different inflections of a word whereas lemmatization removes inflectional endings to return a word to its root morphologically.

**Object standardization** is the process of converting shorthand forms or variant spellings to the formal spelling (e.g., luv to love). One solution is to use a comprehensive dictionary and all misspellings and abbreviations are considered noise and removed.

#### 2.2.2 Text to Features

**Syntactical Parsing** is commonly broken into two types of analysis: dependency grammar and part of speech tagging (PoS). They are used to model out the structure of sentences. Dependency is in the context of grammar. Each sentence is broken down into a triplet relation (i.e., relation, governor and dependent). One can think of the triplet relation in grammar terms as the subject, the verb and the direct (or indirect) object in a sentence. Typically a tree is used to represent the triplet graphically. The main idea of syntactical parsing is to take into account sentence structures before processing a text.

Another common tool to understand syntax is PoS. For instance, many English words can take on different parts-of-speech depending on context (e.g., book – book a flight or reading a book). By having PoS taggings, NLP has additional information to process and understand words from the PoS dimension.

**Statistical features** are the numerical results of converting a text into quantifiable characteristics. Common examples are a count of words, sentences or syllables from a text. From there, one can derive, for instance, the sentiment (i.e., the tone) of a text by counting the proportion of negative words in that text or the readability of a text (e.g., Gunning Fog Index) by calculating the average word count per sentence and the proportion of polysyllabic words in that text.

### Word Embedding

A more advanced statistical feature is **word embedding**, a method of using a vector of numbers to capture different dimensions of a word. Intuitively, one can think of each dimension (i.e., number) as a usage for that word, but in actual implementation the dimensions aren't known.

Exhibit 1 provides a simple illustration of word embedding. The far left column contains the different dimensions of each word that we are trying to understand. The numbers are numerical representations of each of the dimensions that are generated from a training data set. For example, the value of each cell in our illustration is bounded between zero and one inclusively where the magnitude of the values shows the strength of the relationship between the word and the dimension (e.g., Queen and Royalty are extremely closely related by the value 0.990 whereas King and Femininity isn't closely related as the value of 0.050 indicates). As one can imagine if we created hundreds of dimensions for each word, each number in the vector effectively captures one contextual usage of that word.

**Exhibit 1: Simple Example of Word Embedding**

	Words			
Dimensions	King	Queen	Woman	Princess
Royalty	0.990	0.990	0.020	0.980
Masculinity	0.990	0.050	0.010	0.020
Femininity	0.050	0.930	0.999	0.960
Age	0.700	0.600	0.500	0.100

Note: values are fictitious for illustration purposes and the dimensions are unknown in reality

Source: S&P Global Market Intelligence Quantamental Research. Drozd et al., (2016)<sup>6</sup>.

As the number of dimensions increases, word embedding would effectively map out all the different contexts that a word can be used with numerical values indicating the strength of

<sup>6</sup> Drozd, A., Gladkova, A., Matsuoka, S. "Word Embeddings, Analogies, and Machine Learning: Beyond King – Man + Woman = Queen". Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3519–3530

that particular association. For instance by applying vector math to our example, word embedding has learned the concept of 'gender' without us explicitly teaching it that concept (e.g., when you do the vector math: king - masculinity + femininity = queen). Word embedding is one of the fastest growing, cutting-edge techniques where it has been the main tool for a number of major breakthroughs in the NLP space.<sup>7</sup>

### 2.2.3 Testing, Refinement & Assessment of Efficacy

Testing & refinement is the final step in the process. Testing refers to whether NLP has performed a task well using a threshold (e.g., 60% of actual incoming spam emails are categorized as spam emails) that is pre-defined by the user. Refinement is the repeated calibration of the NLP algorithm until it meets the modeler's efficacy threshold. In order to mitigate the risk of data mining, one usually takes the entire input data set and divides it into at least three parts. For instance, fifty percent of the data (perhaps randomly chosen) are designated as the training set and the other fifty percent is set aside as the evaluation set. The training set is then further divided into two equal halves where one of which (basically 25% of the overall data) is labeled as the development set and the other (the other 25%) is designated as the development test set.

$$\begin{aligned} \text{Total data set (100 \%)} &= \text{Development Set (50\%)} + \text{Evaluation Set (50\%)} \\ \text{Development Set (50\%)} &= \text{Training Set (25\%)} + \text{Development Testing Set (25\%)} \end{aligned}$$

The idea is that a model is calibrated on the development training set and the testing of that model's efficacy is done on the development test set. Once the modeler is happy with all the refinements then she takes her model (where she no longer makes any changes at least for the time being) to the evaluation set for the final efficacy testing. For those who are familiar with empirical research, the concept is the same as setting aside in- and out-of-sample data sets. Now using our earlier spam email illustration, the modeler validates her algorithm by looking at the emails that are sorted to the spam email folder. Her algorithm is deemed successful if its success rate in the evaluation set is equal to or greater than her pre-defined threshold for success of 60%.

## 3. Why is NLP Important?

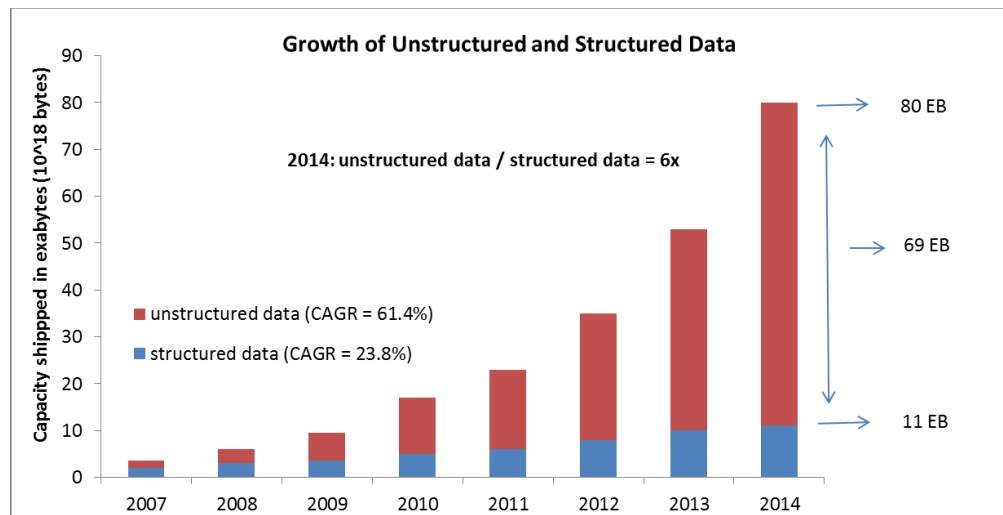
Approximately two and a half exabytes ( $10^{18}$ ) of unstructured data are created daily on the internet right now.<sup>8</sup> In fact to put things into perspective, the amount of unstructured data that was created in the past two days since you have read this primer is equivalent to the same amount of data that was created from the beginning of humankind through the end of

<sup>7</sup> Mikolov T., Yih W., Zweig G. "Linguistic Regularities in Continuous Space Word Representations." Proceedings of NAACL-HLT (2013), 746–751.

<sup>8</sup> Khoso, M. (2016, May 13). How Much Data is Produced Every Day? Retrieved from <http://www.northeastern.edu>.

2003.<sup>9</sup> According to International Data Corporation (IDC), eighty percent of all available data over the internet are unstructured and the growth rate gap between unstructured and structured data is only widening (Exhibit 2). Humans have two ways to process, understand and harness the informational content of these vast, relatively untapped content sets – manual processing, which is infeasible, or automatic processing, where NLP comes into play.

**Exhibit 2: Historical Growth of Unstructured & Structured Data**



Source: S&P Global Market Intelligence Quantamental Research. IDC. 2015.

#### 4. How Can NLP Help Me? – Insights from Earnings Calls

In this section, we illustrate several NLP insights from use cases stemming from S&P Global Market Intelligence's earnings call transcripts:

- Use Case 1: An analysis of the historical relationship between a stock's sentiment changes vs. its forward returns
- Use Case 2: A heat-map of industry-level sentiment trends
- Use Case 3: Language complexity of earnings calls and sell-side analyst selectivity went hand-in-hand in Q2 2017, when managers wanted to soften bad news.
- Use Case 4: Firms with the highest sell-side analyst selectivity ratio underperformed by 2.14% in Q2 2017. This non-NLP use case highlights the speaker type segmentation feature of earnings call transcripts.

<sup>9</sup> DeAngelis, S. F. (2014, Feb.). The Growing Importance of Natural Language Processing [Web log post]. Retrieved Aug 8, 2017, from <http://www.wired.com>

#### 4.1 Details on Loughran and McDonald (2011) Financial Dictionary

In use cases 1 and 2, sentiment changes are measured at the stock-level and at the industry group-level. How is sentiment defined exactly? There are many ways to define sentiment. We use a bag-of-words approach where the sentiment word lists are from the Loughran and McDonald (2011) financial dictionary. Their dictionary has become the de facto financial dictionary for NLP analysis due to its **accessibility**, its **comprehensiveness**, its **financial-specific context**, its **lack of dependency on the transitory nature of its words** and, lastly and perhaps most importantly, its **unambiguous and singularly connoted words**. Details below.

- **Accessibility** – their word lists are readily accessible because they are freely posted online.
- **Comprehensiveness** – the dictionary is comprehensive such that it is difficult for managers to game the system (i.e., circumvent certain words that have empirically been shown to lead to future stock underperformance) because they start with every conceivable English word with all inflections of a word, totaling 80,000+ distinct words in the master word list.
- **Financial-specific context** - they filter their initial master word list down to their sentiment word lists by examining 10-K filings between 1994 and 2008 inclusively.
- **Permanence of words** - their master and sentiment word lists are less transitory because they start with the most comprehensive list of English words possible and, more importantly, the master word list doesn't rely on transitory terms such as iphone.
- **Unambiguous and Singularly Connoted Words** - they arrive at their sentiment word lists containing unambiguous and singularly connoted words by looking at the most frequently occurring words in the 10-Ks from the master word list. From there, they went word-by-word and assessed each of the word's meaning in a business context. At the end of their process, the words that ended up in their word lists are less ambiguous in their meaning with singular connotation.

Their three most important lists of words for our use cases are the master word list, positive and negative sentiment word lists with distinct word counts of 80,000+, 350+ and 2300+, respectively. Examples of positive words are able, abundance, acclaimed, accomplish and so forth. Examples of negative words are abandon, abdicate, aberrant, abetting and so forth.

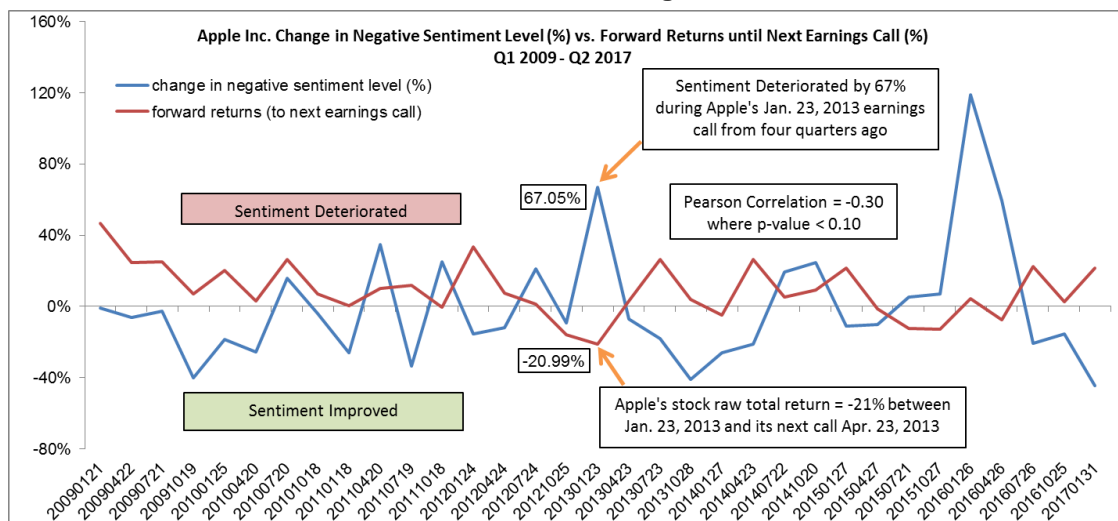
#### 4.2 Stock-Level Sentiment Trends

The first use case is to capture the historical relationship of Apple Inc.'s sentiment level changes from its earnings calls and its forward returns until the next call (Exhibit 3). The changes in sentiment are defined as quarter-over-quarter (QoQ) changes from four quarters ago (to account for seasonality). The sentiment of each of Apple's earnings calls is defined



by the proportion of negative words in its earnings call transcript where the classification of both the negative and the master word list is based on the [Loughran and McDonald \(2011\)](#) financial dictionary. Because sentiment in this use case is measured with negative words, positive (negative) changes reflect sentiment deterioration (improvement). Apple's forward returns until its future calls have been shifted back a quarter such that its sentiment changes and its forward returns are aligned vertically in the Exhibit. One promising observation is that the Pearson correlation is about -0.30 since Q1 2009, which suggests that Apple's forward returns historically go down when its sentiment deteriorates.

**Exhibit 3: Stock Level – Sentiment Change vs. Forward Returns**



Note: Sentiment is defined as the proportion of negative words in an earnings call using [Loughran and McDonald \(2011\)](#). Sentiment changes are measured quarter-over-quarter from four quarters ago. Source: S&P Global Market Intelligence Quantamental Research. Data as of 08/08/2017.

### 4.3 Industry-Level Use Case: Trends in Sentiment Changes

The second use case provides a heat map showing sentiment trends for S&P 500 GICS industry groups. Exhibit 4 shows quarterly sentiment changes for 24 GICS industry groups by calendar quarter between Q3 2016 and Q2 2017 inclusively. The industry groups are sorted by their sentiment changes from Q2 2017 in descending order. Similar to the Apple's example from above, all the sentiment values are QoQ changes from four quarters ago where each industry group's sentiment is aggregated, on an equal-weighted basis, from the stock-level where the sentiment is measured using the proportion of negative words in a stock's earnings call. In order to make the interpretation more intuitive, we multiplied the values by negative one so green (red) values denote improvement (deterioration) in sentiment for the industry groups and the different color shades reflect the magnitude of sentiment changes.

One could easily visualize sentiment trends for an industry group and spot potential inflection points and accelerations. For example, the sentiment improved substantially for banks between calendar quarter Q4 2016 and Q1 2017. Investors would notice this as an

inflection point to the upside and could potentially use the insight as an additional piece of information in their investment decision making process.

**Exhibit 4: S&P 500 Trends in Sentiment Change**

GICS Industry Groups	Calendar Quarters			
	Q3 2016	Q4 2016	Q1 2017	Q2 2017
Consumer Services	-4.1%	-4.0%	16.9%	20.1%
Software & Services	-0.2%	4.1%	-3.9%	19.2%
Transportation	-3.6%	-2.8%	14.6%	19.1%
Diversified Financials	5.4%	13.8%	16.5%	18.0%
Technology Hardware & Equipment	4.8%	7.1%	18.1%	17.5%
Banks	-3.7%	-0.4%	18.1%	16.8%
Capital Goods	-4.9%	9.1%	16.9%	15.5%
Commercial & Professional Services	7.6%	-4.4%	-6.2%	14.6%
Utilities	-4.6%	7.2%	7.0%	14.1%
Insurance	0.0%	14.4%	10.2%	11.9%
Energy	0.1%	13.6%	25.8%	10.3%
Real Estate	-11.5%	-2.4%	0.5%	5.5%
Media	-11.6%	-12.6%	5.0%	4.2%
Materials	5.7%	-1.8%	12.2%	1.4%
Semiconductors & Semiconductor Equipment	6.3%	-3.7%	14.5%	0.5%
Retailing	-18.4%	-0.3%	2.4%	-1.6%
Consumer Durables & Apparel	10.7%	0.2%	-4.7%	-2.6%
Pharmaceuticals, Biotechnology & Life Sciences	-5.7%	-3.6%	-1.3%	-2.9%
Household & Personal Products	-3.3%	-2.9%	-7.8%	-3.9%
Food & Staples Retailing	17.7%	-0.8%	-16.5%	-4.0%
Food, Beverage & Tobacco	2.2%	0.4%	-3.6%	-5.0%
Health Care Equipment & Services	-8.2%	-2.7%	-1.0%	-7.5%
Autos & Components	-14.6%	-4.3%	34.2%	-8.5%
Telecommunication Services	9.1%	-13.6%	-36.9%	-29.4%

Note: Sentiment is defined as the proportion of negative words in an earnings call using [Loughran and McDonald \(2011\)](#). Sentiment changes are measured quarter-over-quarter from four quarters ago where the values are multiplied by -1 to make results easier to interpret. Industry group level values are rolled up equal-weighted from the stock-level. Source: S&P Global Market Intelligence Quantamental Research. Data as of 08/08/2017.

#### 4.4 Does Language Complexity of Earnings Calls and the Sell-Side Analyst Selectivity Ratio Go Hand-in-Hand?

The third use case showcases the segmentation of S&P Global Market Intelligence's earnings call transcripts by speaker types specifically looking at sell-side analysts. The case demonstrates whether managers who have bad news may be attempting to soften that news. We examined the following two dimensions: the language complexity of earnings calls and the selectivity of sell-side analysts who are picked to ask questions during earnings calls in the form of a ratio.

##### 4.4.1 Defining Gunning Fog Index & Our Hypothesis

The language complexity of earnings calls is measured using the (Gunning) fog index (see the y-axis of Exhibit 5), which measures the number of years of formal education that one needs to understand the diction used in the analyzed text (e.g., 16 is equivalent to someone who has completed an undergraduate degree), in our case an earnings call transcript. The fog index has two inputs: the average number of words per sentence and the proportion of

polysyllabic words (threshold is 3 syllables or higher). A higher number of words per sentence and/or a higher proportion of polysyllabic words in a text would result in an increase in the fog index.

Why might a fog index be a good proxy to assess whether managers are trying to soften bad news? First, we surmise that when news is bad (e.g., missed earnings) managers need to disclose it due to their legal or fiduciary obligations. Hoping to mitigate the effect of the news on their stock prices, they may provide long-winded 'explanations' of the news, whereas in the quarters with good news answers to sell-side analysts tend to be shorter and more direct. Secondly, managers may give more long-winded answers to drain the time remaining in the Q&A section of an earnings call where the historical average duration of earnings calls is about an hour.

#### 4.4.2 Defining Sell-Side Analyst Selectivity Ratio and Our Hypotheses

The sell-side analyst selectivity ratio is defined as the percent of the active sell-side coverage that are allowed to ask questions during an earnings call. For instance, Apple Inc. has forty active sell-side coverages prior to one of its earnings calls and its management allowed 10 of the analysts to ask questions, which translates to an analyst selectivity ratio of 25% and in our narrative is interpreted as a high analyst selectivity ratio and viewed negatively.

Our hypothesis is that when a firm has good financial results, it wants everyone to know, especially sell-side analysts because they are great messengers to buy-side money managers and traders. Answers to sell-side questions tend to be shorter and more direct when firms are doing well, which in turn enables managers to allow more analyst questions. When financial results are less favorable, the percent of sell-side analysts who get to ask questions declines due to two possible reasons. One is that managers tend to have lengthier explanations as to why the negative results are perhaps transitory, which consumes additional time that would otherwise be allocated to taking additional analyst questions. Secondly, managers may take multiple questions from analysts who may have a positive view on their firms.

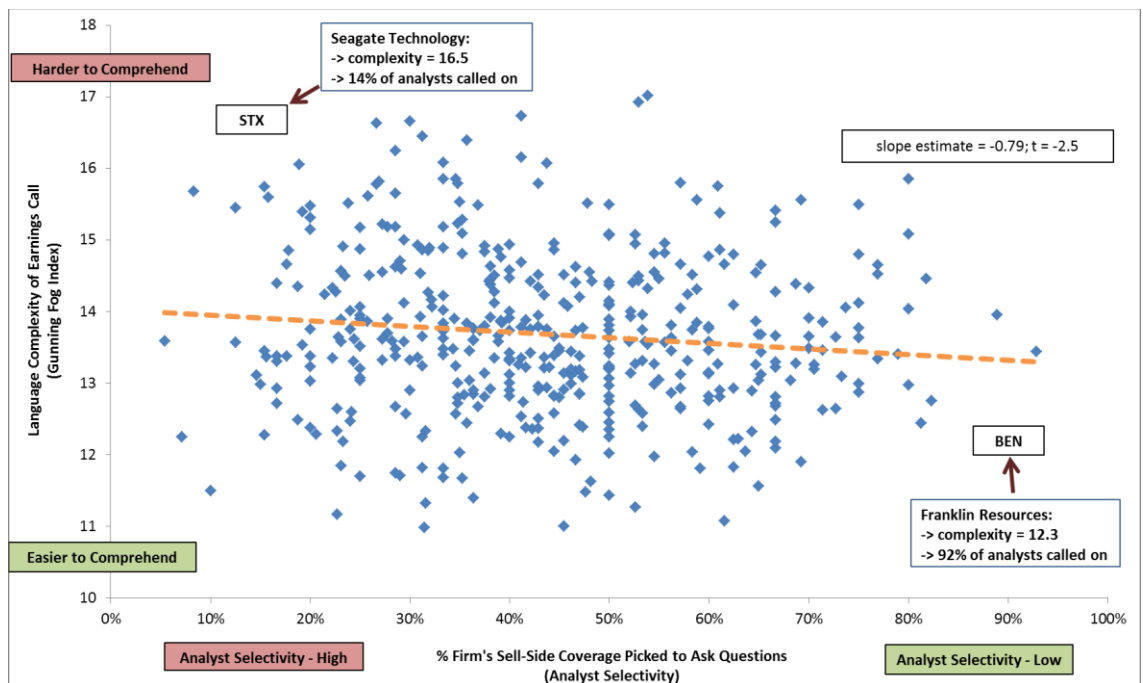
Naturally, there are other ways of measuring analyst selectivity ratio (or language complexity). For instance, one potential issue with this analyst selectivity ratio is that it may have a market-capitalization bias. If a firm has a great number of sell-side analysts covering it, it is generally harder for a firm to call on every analyst who has an active coverage (e.g., Apple Inc. called only 7 of the 43 analysts during its calendar quarter Q2 2017). Because generally larger market-caps have a greater number of sell-side coverage, larger (smaller) market-cap firms may always have a higher (lower) analyst selectivity ratio. In calendar Q2 2017 using a snapshot of market-caps on Mar. 31, 2017, **the spearman correlation**

between each firm's analyst selectivity ratio and its market-cap is **-0.14**.<sup>10</sup> The -0.14 correlation result suggests that there is a relationship between a firm's analyst selectivity ratio and its market-cap (i.e., firms with larger market-caps have lower analyst selectivity ratios), but it isn't very strong. Thus, we are using this flavor of analyst selectivity ratio for the use case without further consideration, mainly because we want to keep the narrative as simple as possible.

#### 4.4.3 Results & Interpretation

In use case 3 (Exhibit 5), our results in Q2 2017 do seem to indicate that firms with managers whose language complexity during earnings calls is higher also took a smaller proportion of questions from their active sell-side coverage. This relationship is captured statistically by the best-fitted, downward sloping orange-dotted line with a t-statistic of about -2.5<sup>11</sup>.

**Exhibit 5: Language Complexity of Earnings Call versus Sell-Side Analyst Selectivity Ratio S&P 500 – Q2 2017**



Source: S&P Global Market Intelligence Quantamental Research. Data as at 08/08/2017

#### 4.5 Firms with the Highest Sell-Side Analyst Selectivity Underperformed in Q2 2017

In our final use case 4, we continue the narrative from use case 3 where we examine the forward returns of firms whose managers called on a smaller percentage of their active sell-

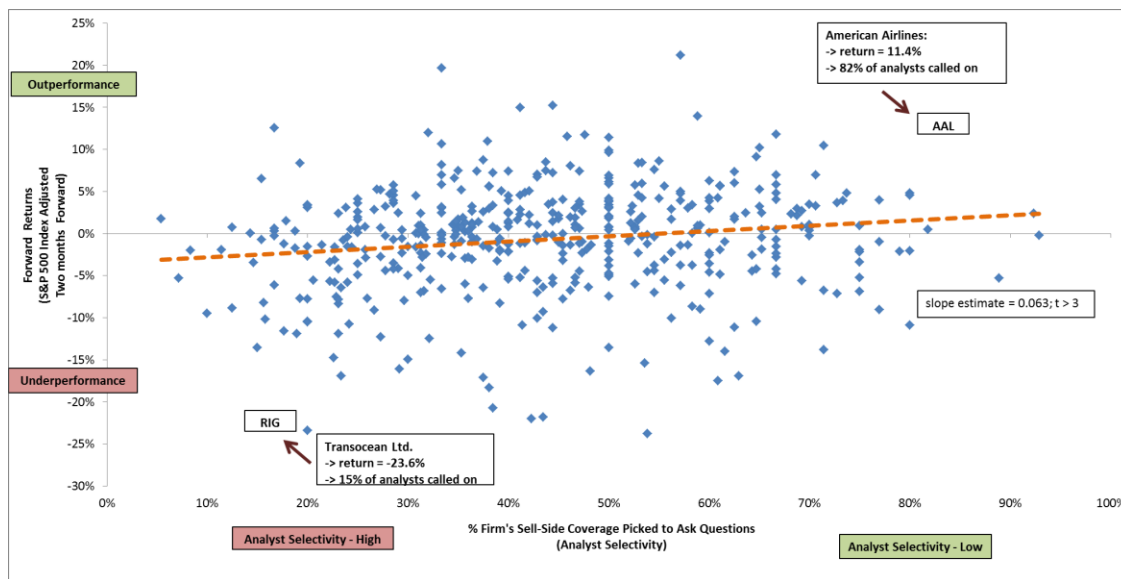
<sup>10</sup> The spearman correlation between each firm's number of active analyst coverage and its market-cap is 0.42. The spearman correlation between each firm's number of active analyst coverage and its analyst selectivity ratio is -0.37.

<sup>11</sup> We are 99% confident statistically that in Q2 2017 there is a positive relationship between the two proxies and the relationship isn't due to random chance.

side coverage. We took two-month forward returns of firms in the S&P 500 Index after their earnings call and plotted against their level of analyst selectivity (Exhibit 6).

We see in Q2 2017 that firms whose managers were more selective underperformed by 2.14% in the ensuing two months (Exhibit 6). The returns are adjusted for S&P 500 returns (assuming a CAPM beta of one) and are calculated using an event study framework where in hindsight we sorted all the firms into tertile bins and went long (short) the firms with the most (least) transparency using analyst selectivity as a proxy. When we fit a line through the scatter plot, the slope is upward sloping as intuition would suggest with a slope estimate of 0.063 with a t-statistic > 3. The result indicates that the underperformance in Q2 isn't due to chance.

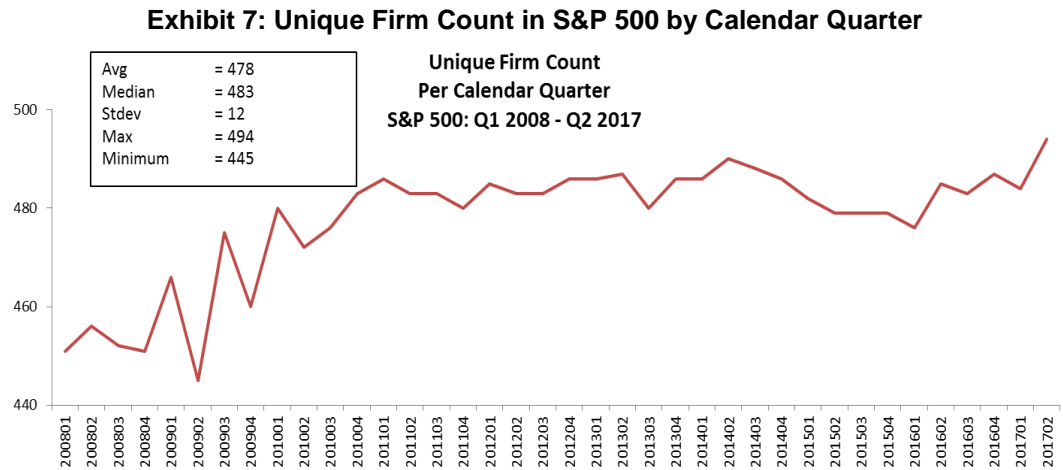
**Exhibit 6: Forward Returns versus Sell-Side Analyst Selectivity Ratio  
S&P 500 – Q2 2017**



Source: S&P Global Market Intelligence Quantamental Research. Data as at 08/08/2017

## 5. Transcripts Data Set

The earnings call transcript data set is a new addition to S&P Global Market Intelligence's Xpressfeed and desktop platforms. The data set starts in calendar quarter Q1 2008 and has very good coverage of all U.S. firms within S&P 500 Index (Exhibit 7). Among its key features, the data set captures the segmentation of earnings calls by speaker types (e.g., managers, sell-side analysts, shareholders etc.) and by sections (e.g., presentation section vs. Q&A section)



Source: S&P Global Market Intelligence Quantamental Research. Data as at 08/28/2017

## 6. Review of Tools & Code Snippets

Analyses for the primer are performed with Python and Matlab. All the NLP is done in Python, a tool that is free, ubiquitous in the field, and robust once the practitioner can navigate the learning curve. The calculations and matrix manipulations are done in Matlab (Python does have a Matlab-like library *numpy*, but we used Matlab due to preference).

### 6.1 Steps to Replicate Examples & Code Snippets

In this section, we walk our readers through our three NLP examples from [Section 4](#) highlighting important sections via code snippets. The entire [code](#) is here. Steps are as follows:

- 
- **Step 1:** Download the Python integrated development environment (IDE). For illustrative purposes, we used Anaconda. <https://www.anaconda.com/download/>

- 
- **Step 2:** Obtain sentiment word lists and save them to text files to ingest by the Python code. For illustrative purposes, we used [Loughran and McDonald \(2011\)](#) sentiment word lists ([https://www3.nd.edu/~mcdonald/Word\\_Lists.html](https://www3.nd.edu/~mcdonald/Word_Lists.html)) to identify the sentiment of each earnings call transcript.

- 
- **Step 3:** Format earnings call transcripts in the following way. There are five columns in the input files where the columns going from left to right are: stock identifier, earning call date, hour and minute, sequence identifier of every earnings call and the content of every earnings call
- 

- **Step 4:** Load Python libraries that one needs (lines 22 – 35)

```
#####
import os # import the os module
import pandas as pd # using pandas to create data type 'DataFrame' use
import time # time module can be used now e.g., time.time()
import nltk

#####
fpath = 'C:/Users/fzhao/Desktop/NLP/rawData'
os.chdir(fpath) # chg working directory to fpath
```

---

- **Step 5:** Load in Loughran and McDonald (2011) word lists (lines 43 - 69).

```
fn = 'wordLists\lmMasterList.txt' # input file
tmpWordListPandaObj = pd.read_csv(fn, sep='\t', header=None) # read in tab-delimit
tmpWordListPandaObj.columns = ['master'] # rename cols
master_list = list(tmpWordListPandaObj.master)
master_list = [itr for itr in master_list if not isinstance(itr,float) ]
master_list = [itr.lower() for itr in master_list] # make all lowercase
```

---

- **Step 6:** Define the functions that you are going to use. The one function included below outputs the frequency count of words (lines 84 - 160).

```
def _out_word_freq1(list1, list2):
    wordFreqDictObj = {}
    for item in list2:
        if item in list1:
            wordFreqDictObj[item] = wordFreqDictObj.get(item, 0) + 1
    return wordFreqDictObj
```

---

- **Step 7:** Have a control structure to traverse through time periods and firms. We decided to traverse one quarter at a time across all firms for that quarter. We used the **while** loop control structure whereas one could also use the **for loop** control structure. Relative to other languages like Matlab, C++, etc., the **for loop** in Python is very robust and powerful, which also makes some **for loop** syntaxes unintuitive. If you decide to use the **for loop** instead, perhaps consider using the key word **enumerate** where you can actually observe and utilize the iterator in the **for loop** (lines 173 - 213).

```

start_time                = time.time(); # record time when a piece of code is executed
yrItr                     = 0
while yrItr < len(yrVec):

    qtrItr                 = 0
    while qtrItr < len(qtrVec):

        #creating filename to read in
        fn                 = str('transcriptsAll\\sp9_q' + str(qtrVec[qtrItr]) +

        #print(yrVec[yrItr], qtrVec[qtrItr])

        #check file exists
        if os.path.isfile(fn) and os.stat(fn).st_size != 0:

            tmpPandaObj     = pd.read_csv(fn, sep='\t', header=None) # read
            tmpPandaObj.columns = ['stockId','dt','hmm','seq','ecalls'] # read

```

- **Step 8:** This step is about tokenizing and preprocessing of an earnings call. In the screenshot below, we use the non-native, but well-known, **nlTK** NLP library and specifically the method **word\_tokenize** to parse an earnings call into tokens where the function parses whenever it encounters a (user-defined) punctuation or white space. We also show other text preprocessing that we did: i) removal of punctuation ii) removal of empty spaces between words and iii) standardizing all text to lower case and so forth (lines 234 - 240).

```

#####
# nltk mthds
tokensListObj             = nltk.word_tokenize(dsRawSeriesObj) # tokenize all words
tokensListObj             = [itr.lower() for itr in tokensListObj] # make all lowercase

# punct
punct                    = string.punctuation # native punctutation string
tokensListObj            = ["!\"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~"; # string.punctuation less '-'
tokensListObj            = ["".join([j for j in i if j not in punct]) for i in tokensListObj]
tokensListObj            = [itr for itr in tokensListObj if itr] # keep on concat if item in list isnt empty

```



- 
- **Step 9:** Count the frequency of total words and negative words as defined by the [Loughran and McDonald \(2011\)](#) dictionary, utilizing our user-defined functions. First, find the intersection of the vector of words that belong to both an earnings call and the master word list. The function only outputs a unique vector of words. So we feed the vector of overlapping words and the earnings call into our user-defined function earlier to output a hash table where the words are the keys and their frequencies of usage as another column. Lastly, traverse through all firms in a calendar quarter and output each computation to the data structure **panda**. Then you can just divide the number of negative words in each transcript by the total number of word in that transcript to calculate a negative sentiment level and from there calculate the change in the sentiment level between quarters once you output the rest of the data set (lines 246 - 255).

```

tmpWordsList          = list(set(tokensListObj) & set(master_list))
tokensNmasterList     = tmpWordsList
# tmpWordsList2      = set

if not tmpWordsList:
    tmpInt             = 0
else:
    tmpDfObj           = _out_word_freq1(tmpWordsList, tokensListObj)
    masterWordsDsDictObj = tmpDfObj
    tmpInt             = sum(tmpDfObj.values())
    masterCnt          = tmpInt

```

## References

Antweiler, W., Frank, M. Z., 2004. "Is all that talk just noise? the information content of internet stock message boards." *Journal of Finance* 59, 1259 -1293.

Bansal, S. (2017, Jan. 12). Ultimate Guide to Understand & Implement Natural Language Processing (with codes in Python) [Web log post]. Retrieved Aug. 8, 2017, from <https://www.analyticsvidhya.com>

Bradley, M., Lang, P., 1999. "Active norms for English words (ANEW): Stimuli, instruction manual and active ratings." Technical report C-1, The Center for Research in Psychophysiology, University of Florida.

Copeland, M. (2016, Jul. 29). What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning? [Web log post]. Retrieved Aug. 8, 2017, from <https://blogs.nvidia.com>

DeAngelis, S. F. (2014, Feb.). The Growing Importance of Natural Language Processing [Web log post]. Retrieved Aug 8, 2017, from <http://www.wired.com>

Droz, A., Gladkova, A., Matsuoka, S. "Word Embeddings, Analogies, and Machine Learning: Beyond King – Man + Woman = Queen". Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3519–3530

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. "The incremental information content of tone change in management discussion and analysis." *Review of Accounting Studies* 15, 915-953.

Griffin, P., 2003. "Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings." *Review of Accounting Studies* 8, 433-460.

Hanley, K. W., Hoberg, G., 2010. "The information content of IPO prospectuses." *Review of Financial Studies* 23, 2821-2864.

Khoso, M. (2016, May 13). How Much Data is Produced Every Day? Retrieved from <http://www.northeastern.edu>

LaFrance A. (2017, June 15). An Artificial Intelligence Developed Its Own Non-Human Language. <https://www.theatlantic.com>.

Loughran, T., AND B. McDonald. "When is a Liability not a Liability? Textual analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (2011): 35-65.

Loughran, T., AND B. McDonald. "IPO First-day Returns, Offer Price Revisions, Volatility, and Form S-1 Language." *Journal of Financial Economics* 109 (2013): 307-326.

Loughran, T., AND B. McDonald. "Measuring Readability in Financial Disclosures." *Journal of Finance* 69 (2014): 1643-1671.

Loughran, T., AND B. McDonald. "The Use of Word Lists in Textual Analysis." *Journal of Behavioral Finance* 16 (2015): 1-11.

Mikolov T., Yih W., Zweig G. "Linguistic Regularities in Continuous Space Word Representations." *Proceedings of NAACL-HLT* (2013), 746–751.

Mozur, P. (2017, May 25). Google's A.I. Program Rattles Chinese Go Master as It Wins Match. Retrieved from <http://www.nytimes.com>.

Turing A.M. "Computing Machinery and Intelligence." *Mind* 49 (1950), 433-460

## Our Recent Research

### **July 2017: Natural Language Processing Literature Survey**

In client conversations, Natural Language Processing (NLP) and the analysis of unstructured data is a topic of regular conversation. S&P Global Market Intelligence offers several unstructured datasets garnering market attention. The first is earnings call transcripts, with unique speaker id's to identify who is speaking on the call. The second data set is the text content in the 10-K. In advance of a publication of Quantamental primer on NLP next month which will take readers through the process of handling unstructured data and generating sentiment scores, we offer this literature survey. What follows are ten papers that the team has identified as being of particular interest to investors on this topic.

### **June 2017: Research Brief: Four Important Things to Know About Banks in a Rising Rate Environment**

With the Fed signaling further rate hikes ahead, bank investors may want to know which investment strategies have worked best in a rising rate environment historically. This paper leverages our empirical work on the SNL Bank fundamental data to aid investors in selecting bank stocks as rates rise.

### **April 2017: Banking on Alpha: Uncovering Investing Signals Using SNL Bank Data**

This study leverages S&P Global Market Intelligence's SNL Financial data to answer three questions of importance to bank investors: 1. Which widely-used investment strategies have historically been profitable? 2. Which lesser-known strategies deserve wider attention? 3. How do these strategies perform across varying macro environments: rising vs. falling interest rates and above- vs. below-average financial stress?

### **March 2017: Capital Market Implications of Spinoffs**

Spinoff activities have picked up in recent years. In 2015, more than \$250 billion worth of spinoff transactions were closed globally - the highest level in the last 20 years. This report analyzes the short- and long-term performance of spun-off entities and their parent companies in the U.S. and international markets. We also examine a related but distinct corporate restructuring activity – equity carve-outs, which separate a subsidiary through a public offering.

### **January 2017: U.S. Stock Selection Model Performance Review 2016**

2016 proved to be a challenging year for active investing. Against a backdrop of a sharp selloff in equities at the beginning of the year and political uncertainty over the course of the year, valuation was the only fundamental investing style that delivered positive excess returns. In this report, we review the performance of S&P Global Market Intelligence's four U.S. stock selection models in 2016.

**November 2016: Electrify Stock Returns in U.S. Utilities**

The U.S. utilities sector has performed especially well in the past several years as the Federal Reserve and central banks around the world enacted accommodative monetary policies to spur growth. As global active investors flock to the U.S. utilities sector in search of yields and high risk-adjusted returns, we explore a number of utility-specific metrics from a unique database that is dedicated to the utilities sector – S&P Global Market Intelligence’s Energy (Source: SNL Energy) – to ascertain whether investors could have historically made stock selection decisions within the sector to achieve excess returns.

**October 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 2**

SNL Financial’s (“SNL”) 1 global real estate database contains property level and geographical market-based demographic information that can be difficult for investors to obtain. These unique data points are valuable to investors seeking an understanding of the relationship between property level information and future stock price movement. In this report, we demonstrate how investors can use these data points as alpha strategies. Our back-tests suggest that metrics constructed from property level information may provide insights about future price direction not captured by fundamental or estimates data. Investors may want to consider incorporating information on a REIT’s property portfolio when building a robust REIT strategy

**September 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 1**

This month REITs (Real Estate Investment Trusts) have been separated from the GICS (Global Industry Classification Standard) Financial sector into a sector of their own. Even prior to the sector reclassification, investors have been attracted to REITs’ strong performance and attractive yield. REITs differ from traditional companies in several important ways. Metrics that investors typically use to value or evaluate the attractiveness of stocks such as earnings yield or book-to-price are less meaningful for REITs. For active investors interested in understanding their REITs portfolio, an understanding of the relationship between REIT financial ratios and price appreciation is instructive. Is dividend yield relevant? What about funds from operations (“FFO”), one of the most widely used metrics?

**August 2016: Mergers & Acquisitions: The Good, the Bad and the Ugly (and how to tell them apart)**

In this study we show that, among Russell 3000 firms with acquisitions greater than 5% of acquirer enterprise value, post-M&A acquirer returns have underperformed peers in general. Specifically, we find that:

- Acquirers lag industry peers on a variety of fundamental metrics for an extended period following an acquisition.

- Stock deals significantly underperform cash deals. Acquirers using the highest percentage of stock underperform industry peers by 3.3% one year post-close and by 8.1% after three years.
- Acquirers that grow quickly pre-acquisition often underperform post-acquisition.
- Excess cash on the balance sheet is detrimental for M&A, possibly due to a lack of discipline in deploying that cash.

**July 2016: Preparing for a Slide in Oil Prices -- History May Be Your Guide**

With the price of West Texas Intermediate (WTI) in the mid-forties, oversupply concerns and the continued threat of a global slowdown have led many to fear a resumed oil price decline. The year-to-date performance of Oil & Gas (O&G) companies, particularly Integrated O&G entities has been strong, further contributing to concerns that oil may be poised to retrench.

**June 2016: Social Media and Stock Returns: Is There Value in Cyberspace?**

This review of social media literature represents a selection of articles we found particularly pragmatic and/or interesting. Although we have not done research in the area of social media, we are always on the hunt for interesting insights, and offer these papers for your thoughtful consideration.

**April 2016: An IQ Test for the “Smart Money” – Is the Reputation of Institutional Investors Warranted?**

This report explores four classes of stock selection signals associated with institutional ownership ('IO'): Ownership Level, Ownership Breadth, Change in Ownership Level and Ownership Dynamics. It then segments these signals by classes of institutions: Hedge Funds, Mutual Funds, Pension Funds, Banks and Insurance Companies. The study confirms many of the findings from earlier work – not only in the U.S., but also in a much broader geographic scope – that Institutional Ownership may have an impact on stock prices. The analysis then builds upon existing literature by further exploring the benefit of blending 'IO' signals with traditional fundamental based stock selection signals.

**March 2016: Stock-Level Liquidity – Alpha or Risk? - Stocks with Rising Liquidity Outperform Globally**

Most investors do not associate stock-level liquidity as a stock selection signal, but as a measure of how easily a trade can be executed without incurring a large transaction cost or adverse price impact. Inspired by recent literature, such as Bali, Peng, Shen and Tang (2012), we show globally that a strategy of buying stocks with the highest one-year change in stock-level turnover has historically outperformed the market and has outperformed strategies of buying stocks with strong price momentum, attractive valuation, or high quality. One-year change in stock-level turnover has a low correlation (i.e., <0.15) with commonly used stock selection signals. When it is combined with these signals, the composites have yielded higher excess returns and information ratios (IR) than the standalone raw signals.

**February 2016: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2015**

Since the launch of the four S&P Capital IQ® U.S. stock selection models in January 2011, **the performance of all four models (Growth Benchmark Model, Value Benchmark Model, Quality Model, and Price Momentum Model) has been positive each year.** The models' key differentiators – a distinct formulation for large cap versus small cap stocks, incorporation of industry specific information for the financial sector, sector neutrality to target stock specific alpha, and factor diversity – enabled the models to outperform across disparate market environments. In this report, we assess the underlying drivers of each model's performance in 2015 and since inception (2011), and provide full model performance history from January 1987.

**January 2016: What Does Earnings Guidance Tell Us? – Listen When Management Announces Good News**

This study examines stock price movements surrounding earnings per share (EPS) guidance announcements for U.S. companies between January 2003 and February 2015 using S&P Capital IQ's Estimates database. Companies that experienced positive guidance news, i.e. those that announced optimistic guidance (guidance that is higher than consensus estimates) or revised their guidance upward, yielded positive excess returns. We focus on guidance that is not issued concurrent with earnings releases in order to have a clear understanding of the market impact of guidance disclosures. We also explore practical ways in which investors may benefit from annual and quarterly guidance information.

**December 2015: Equity Market Pulse – Quarterly Equity Market Insights Issue 6**

**November 2015: Late to File - The Costs of Delayed 10-Q and 10-K Company Filings**

**October 2015: Global Country Allocation Strategies**

**September 2015: Equity Market Pulse – Quarterly Equity Market Insights Issue 5**

**September 2015: Research Brief: Building Smart Beta Portfolios**

**September 2015: Research Brief – Airline Industry Factors**

**August 2015: Point-In-Time vs. Lagged Fundamentals – This time i(t)'s different?**

**August 2015: Introducing S&P Capital IQ Stock Selection Model for the Japanese Market**

**July 2015: Research Brief – Liquidity Fragility**

June 2015: Equity Market Pulse – Quarterly Equity Market Insights Issue 4

May 2015: Investing in a World with Increasing Investor Activism

April 2015: Drilling for Alpha in the Oil and Gas Industry – Insights from Industry Specific Data & Company Financials

March 2015: Equity Market Pulse – Quarterly Equity Market Insights Issue 3

February 2015: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2014

January 2015: Research Brief: Global Pension Plans - Are Fully Funded Plans a Relic of the Past?

January 2015: Profitability: Growth-Like Strategy, Value-Like Returns - Profiting from Companies with Large Economic Moats

November 2014: Equity Market Pulse – Quarterly Equity Market Insights Issue 2

October 2014: Lenders Lead, Owners Follow - The Relationship between Credit Indicators and Equity Returns

August 2014: Equity Market Pulse – Quarterly Equity Market Insights Issue 1

July 2014: Factor Insight: Reducing the Downside of a Trend Following Strategy

May 2014: Introducing S&P Capital IQ's Fundamental China A-Share Equity Risk Model

April 2014: Riding the Coattails of Activist Investors Yields Short and Long Term Outperformance

March 2014: Insights from Academic Literature: Corporate Character, Trading Insights, & New Data Sources

February 2014: Obtaining an Edge in Emerging Markets

February 2014: U.S. Stock Selection Model Performance Review

January 2014: Buying Outperformance: Do share repurchase announcements lead to higher returns?



**October 2013: Informative Insider Trading - The Hidden Profits in Corporate Insider Filings**

**September 2013: Beggar Thy Neighbor – Research Brief: Exploring Pension Plans**

**August 2013: Introducing S&P Capital IQ Global Stock Selection Models for Developed Markets: The Foundations of Outperformance**

**July 2013: Inspirational Papers on Innovative Topics: Asset Allocation, Insider Trading & Event Studies**

**June 2013: Supply Chain Interactions Part 2: Companies – Connected Company Returns Examined as Event Signals**

**June 2013: Behind the Asset Growth Anomaly – Over-promising but Under-delivering**

**April 2013: Complicated Firms Made Easy - Using Industry Pure-Plays to Forecast Conglomerate Returns.**

**March 2013: Risk Models That Work When You Need Them - Short Term Risk Model Enhancements**

**March 2013: Follow the Smart Money - Riding the Coattails of Activist Investors**

**February 2013: Stock Selection Model Performance Review: Assessing the Drivers of Performance in 2012**

**January 2013: Research Brief: Exploiting the January Effect Examining Variations in Trend Following Strategies**

**December 2012: Do CEO and CFO Departures Matter? - The Signal Content of CEO and CFO Turnover**

**November 2012: 11 Industries, 70 Alpha Signals -The Value of Industry-Specific Metrics**

**October 2012: Introducing S&P Capital IQ's Fundamental Canada Equity Risk Models**

**September 2012: Factor Insight: Earnings Announcement Return – Is A Return Based Surprise Superior to an Earnings Based Surprise?**

**August 2012: Supply Chain Interactions Part 1: Industries Profiting from Lead-Lag Industry Relationships**

July 2012: Releasing S&P Capital IQ's Regional and Updated Global & US Equity Risk Models

June 2012: Riding Industry Momentum – Enhancing the Residual Reversal Factor

May 2012: The Oil & Gas Industry - Drilling for Alpha Using Global Point-in-Time Industry Data

May 2012: Case Study: S&P Capital IQ – The Platform for Investment Decisions

March 2012: Exploring Alpha from the Securities Lending Market – New Alpha Stemming from Improved Data

January 2012: S&P Capital IQ Stock Selection Model Review – Understanding the Drivers of Performance in 2011

January 2012: Intelligent Estimates – A Superior Model of Earnings Surprise

December 2011: Factor Insight – Residual Reversal

November 2011: Research Brief: Return Correlation and Dispersion – All or Nothing

October 2011: The Banking Industry

September 2011: Methods in Dynamic Weighting

September 2011: Research Brief: Return Correlation and Dispersion

July 2011: Research Brief - A Topical Digest of Investment Strategy Insights

June 2011: A Retail Industry Strategy: Does Industry Specific Data tell a different story?

May 2011: Introducing S&P Capital IQ's Global Fundamental Equity Risk Models

May 2011: Topical Papers That Caught Our Interest

April 2011: Can Dividend Policy Changes Yield Alpha?

April 2011: CQA Spring 2011 Conference Notes

March 2011: How Much Alpha is in Preliminary Data?

February 2011: Industry Insights – Biotechnology: FDA Approval Catalyst Strategy

January 2011: [US Stock Selection Models Introduction](#)

January 2011: [Variations on Minimum Variance](#)

January 2011: [Interesting and Influential Papers We Read in 2010](#)

November 2010: [Is your Bank Under Stress? Introducing our Dynamic Bank Model](#)

October 2010: [Getting the Most from Point-in-Time Data](#)

October 2010: [Another Brick in the Wall: The Historic Failure of Price Momentum](#)

July 2010: [Introducing S&P Capital IQ's Fundamental US Equity Risk Model](#)

Copyright © 2017 by S&P Global Market Intelligence, a division of S&P Global Inc. All rights reserved.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively, S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers, (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON “AS IS” BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT’S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence’s opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global Market Intelligence does not act as a fiduciary or an investment advisor except where registered as such. S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain non-public information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its Web sites, [www.standardandpoors.com](http://www.standardandpoors.com) (free of charge), and [www.ratingsdirect.com](http://www.ratingsdirect.com) and [www.globalcreditportal.com](http://www.globalcreditportal.com) (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at [www.standardandpoors.com/usratingsfees](http://www.standardandpoors.com/usratingsfees).