Machine Invasion: Automation in Information Processing and the Cross-section of Stock Returns

August 2019

Raunaq S. Pungaliya[1]

Yanbo Wang[2]

Abstract

We separate downloads on the SEC EDGAR database into human and machine actions by the intensity of information retrieval (Ryans, 2017). The split shows that the extent of machine downloads has risen 35 times since 2004, accounting for over 96% of total downloads as of 2016. We formally investigate the relationship of machine automation in information processing and the cross-section of stock returns. We find that stocks in the lowest quintile of machine coverage outperform those in the highest quintile by 6 to 7% annually after adjusting for risk. Our results are further supported by a natural experiment on the phase implementation of XBRL tags that enabled machine readable financial disclosure. Our results are consistent with recent theoretical work on: (1) big data (Begenau, Farboodi, and Veldkamp, 2018) since we show that higher machine coverage is associated with lower expected returns, and (2) with Acemoglu and Restrepo (2018) amongst others as we find that machine and human labor are substitutes for the same information processing task, while being complements for sequential and more complex tasks.

Keywords: Machine automation, SEC filings, demand for financial information, big data, stock returns, machine-readable disclosure.

JEL Classification: M41, G1, D8.

[1] Sungkyunkwan University, International Hall #339, 25-2 Sungkyunkwan-ro, Seoul, South Korea, raunaq@skku.edu
[2] Sungkyunkwan University, International Hall #331, 25-2 Sungkyunkwan-ro, Seoul, South Korea, yanbo.wang@skku.edu

Machine Invasion: Automation in Information Processing and the Cross-section of Stock Returns

August 2019

Abstract

We separate downloads on the SEC EDGAR database into human and machine actions by the intensity of information retrieval (Ryans, 2017). The split shows that the extent of machine downloads has risen 35 times since 2004, accounting for over 96% of total downloads as of 2016. We formally investigate the relationship of machine automation in information processing and the cross-section of stock returns. We find that stocks in the lowest quintile of machine coverage outperform those in the highest quintile by 6 to 7% annually after adjusting for risk. Our results are further supported by a natural experiment on the phase implementation of XBRL tags that enabled machine readable financial disclosure. Our results are consistent with recent theoretical work on: (1) big data (Begenau, Farboodi, and Veldkamp, 2018) since we show that higher machine coverage is associated with lower expected returns, and (2) with Acemoglu and Restrepo (2018) amongst others as we find that machine and human labor are substitutes for the same information processing task, while being complements for sequential and more complex tasks.

Keywords: Machine automation, SEC filings, demand for financial information, big data, stock returns, machine-readable disclosure.

JEL Classification: M41, G1, D8.

1. Introduction

"Behind the models employed by quants are algorithms, or investment recipes, that automatically buy and sell based on pre-set inputs. Lately, they're ***dumping stocks****, traders and investors say. The speed and magnitude of the move probably are being* ***exacerbated by machines and model-driven trading****," says Neal Berger, who runs Eagle's View Asset Management, which invests in hedge funds and other vehicles. Human beings tend not to react this fast and violently. (The Wall Street Journal, December 25th, 2018)"*

*"****Automating technologically sleepy parts of financial services*** *could trigger an upheaval for the industry and its employees, analysts say. Citi estimates suggest US and European banks* ***could shed nearly two million jobs in the next decade****, as new technologies transform the sector and new competitors challenge lenders on their traditional turf (Financial Times, May 5th, 2016)"*

The rise of machine automation has raised concerns about whether (1) machine activity has a positive or negative impact on risk in the financial market, and whether (2) automated technologies empower or replace human labor (Acemoglu and Restrepo, 2018; Agrawal, Gans, Goldfarb, 2019). In this paper we shed light not only on the rising impact of machine automation in the financial market, but also on the changing relation between human labor and machines with the rapid emergence of new technologies. We explore these issues by studying the impact of machine automation in information processing on the cross section of stock returns, as well as the conditions under which such impact is amplified.

The previous literature on information market participants has mainly focused on how human activities (e.g. analysts, journalists in the media, and traders) affect the financial market outcome through information processing.[1] This paper is one of the first to empirically examine how non-human actors (i.e. machine automation in information processing) can impact the financial

[1] For example, the role of analysts is studied in Frankel, Kothari, and Weber (2006), the role of media is explored in Tetlock, Saar-Tsechansky, and Macskassy (2008) and Fang and Peress (2009), and the role of retail and institutional traders is documented in Da, Engelberg, and Gao (2011) and Ben-Rephael, Da, Israelsen (2017) respectively.

market. In so doing, our work contributes to the literature as the study of non-human "machine" agents helps to paint a fuller picture of the state of information discovery in financial markets today.

We proxy for machine automation in information processing using the number of machines "covering" a firm for a given month. Our analysis relies on a comprehensive webserver log of IP-address level access requests for each filing in the Security and Exchange Commission's EDGAR financial statement disclosure database. We follow recent work in accounting to classify each individual IP-address into human or machine based on the intensity of information retrieval (Ryans, 2017).[2] Strikingly, we find that the extent of machine attributable downloads has risen 35 times since 2004, accounting for over 96% of all downloads in the EDGAR database in 2016 (Figure 1 A). While it has long been known that machine automation has been increasingly responsible for a dominant fraction of trading activity, our analysis suggests that machine automation also seems to dominate hard information collection from fundamental sources like the SEC's EDGAR database (Zuckerman and Hope, 2018).[3] The rise of automation in information collection is an important facet of today's financial marketplace, but this change has been largely ignored in the empirical literature which has traditionally focused more on human activity.

Has machine automation changed financial markets for the better or worse? First, new studies have argued that the analysis of big data (that machines enable) can help lower the equity premium. In an important paper, Begenau, Farboodi, and Veldkamp (2018) theorize that the availability of big data reduces uncertainty about the future course of events related to the firm, which in turn reduces the risk premium. They also show that the availability of computational

[2] We classify an IP address as a machine if any of the following conditions are met: (1) the daily file downloads from that IP are more than 500, (2) the per-minute download from the IP are more than 25, or (3) the number of firms downloaded per minute from the IP are more than 3. Unlike prior studies that emphasize human downloads as true requests for information (by "throwing away" machine downloads), our study focuses on the rise of machine downloads specifically.

[3] More broadly, 80-90% of trades in the United States are automated leaving only about 10% that can be attributed to direct trades made by "real-life" human actors.

power amplifies the reduction in risk. Their argument suggests that machines act as positive agents as better information processing lowers the conditional variance and covariance of the firm's payoffs increasing their predictability, reducing their uncertainty, and thus lowering the cost of capital for the firm. Specifically, machines achieve this effect by processing and making a large volume of information available at a faster speed compared to humans alone. In other words, machines contribute positively to financial markets by enhancing information discovery and reducing risk premia.

A second opposing view sees the rise of these machine-driven trades through a more cautious lens. This is because unknown "tail risks" in machine automation can bring about unintended consequences and increase the risk of financial instability as seen in the flash crash of 2010 (Kirilenko and Lo, 2013; Kirilenko, Kyle, Samadi, and Tuzun, 2017). These left tail events can act to increase risk premia (Kelly and Jiang, 2014). The share of equity market volume attributed to algorithmic trading in the United States has grown from 20% in 2004 to 85% in 2012 (Glantz and Kissel, 2013). This rise in algorithmic trading share has coincided with the growth in the share of downloads attributed to machines as seen in Figure 1 B. If automation in information collection (machine coverage) is related with algorithmic trading, which in turn exacerbates the risk premium, machine coverage can be associated with an increase in the cost of capital.[4]

Instead of better or worse, a third view posits that human coverage and machine coverage are not different in their implication on financial markets. In other words, this view argues that as machine coverage merely reflects the will of human analysts behind the computer, machine coverage and human coverage should have similar end effects. Veldkamp (2006), Barber and Odean (2007), and Da, Engelberg, and Gao (2011) argue that human attention increases the trading

[4] We show our machine coverage proxy based on observable downloads might also be related to automation in the later stage of the information processing in Section 5. In this context, it is also relevant to note that 29% of the trades in the United States can be attributed to sophisticated quantitative hedge funds that have the means to automate their entire trading process including information retrieval.

demand for both buyer and sellers of an asset, while the selling demand is truncated by the short selling constraint (as it is always easier to buy than to sell short). The net effect is that asset prices can be inflated due to high human attention for a short period, before reversing its course. Thus, this view implies that machine attention should similarly inflate prices in the short term, before they reverse.

What then is the impact of machine automation on financial markets? The three competing theories outlined above provide divergent predictions, motivating the necessity of our empirical study to shed empirical light on this theoretical disagreement. If the first "risk reduction" hypothesis is true, we should see stocks with higher machine coverage exhibit lower returns as a result of lowered risk premia. The second "tail risk" hypothesis has the exact opposite implication compared to the first hypothesis: stocks with higher machine coverage should exhibit higher returns due to increased risk premia. Like the second hypothesis, the third "immateriality" hypothesis also implies that higher machine coverage should exhibit higher returns, however unlike the second hypothesis where the price run-up should be permanent, the third hypothesis implies that the price run-up caused by attention should reverse due to subsequent correction of the mispricing driven by the behavioral bias.

Our novel setting provides us the ability to attribute information requests to machines and humans and thus empirically test these competing theories. We follow seminal work by Hong, Lim, and Stein (2000), Fang and Peress (2009), and Hillert, Jacobs, and Müller (2014) to first examine the determinants of machine coverage. We find that machine coverage is strongly associated with firm size as well as information arrival (e.g. SEC filing and media news arrival). We therefore face the same challenge as Hong, Lim and Stein (2000) who find that analyst coverage is largely correlated to firm size. In their study, they isolate the net effect of analyst coverage by constructing a residual analyst coverage measure after parsing out the contribution of firm size, amongst other firm specific factors. We closely follow the approach of Hong, Lim, and Stein (2000) and Hillert,

Jacobs, and Müller (2014) and estimate adjusted machine coverage as the residual value in a cross-sectional regression of the main determinants of machine coverage, details of which are explained in section 2.3.

We conduct a standard portfolio test and find that stocks in the lowest quintile of (adjusted) machine coverage outperform those in the highest quintile by 6 to 7% annually. The performance difference listed is the annualized value based on monthly alphas (between 51 to 58 basis points) for a variety of risk models including the Fama-French (2015) five-factor model and Daniel-Grinblatt-Titman-Wermer (1997) [DGTW] adjusted returns over our sample period from 2004 to 2016. In addition to portfolio-based tests, we validate our main findings using Fama-Macbeth regressions. These results provide empirical support for recent theoretical work of Begenau, Farboodi, and Veldkamp (2018) as higher machine coverage significantly lowers the expected return on stocks. Importantly, the strong negative correlation that we observe between machine coverage and subsequent returns do not support the "tail risk" and "immateriality" hypotheses.

The rise of machine automation has also drawn the attention of the SEC. In a keynote address in 2018, Scott Bauguess, the Deputy Chief Economist at the Division of Economic and Risk Analysis at the SEC outlined the role of SEC's initiatives on machine readability to "*fuel the machine learning revolution*." Specifically, in 2009, the SEC mandated that financial statements be simultaneously filed using XBRL (eXtensible Business Reporting Language) to improve the transparency of financial disclosure to automated machines in 3-phases over 2009 to 2011 (Bartley et al., 2011). As the presence of XBRL tags for data items and sections makes it easier for machines to parse content in SEC filings, we expect that the introduction of XBRL will reduce information processing costs (Blankespoor, 2019). We use the introduction of XBRL as a natural experiment to validate the relationship between the machine coverage and stock returns. Our difference-in-difference analysis shows that the exogenous increase in the capability of machines to process information due to XBRL is associated with a strengthening of the machine and stock return

relationship. Thus, results from this natural experiment suggest a causal relation between machine coverage and stock returns.

This new machine dominated environment leads us to naturally question the role of human analysis *in the presence of* machine automation. The existence of both machine and human agents in the EDGAR universe offers us a unique controlled environment to study the role of machines versus humans for the same activity or task. Does machine automation act as a substitute or a complement to human labor? For example, Sachs and Kotlikoff (2012) and Nordhaus (2015) argue that automation is capital-augmenting, where automation increases the productivity and thus return to capital. Cheaper capital then replaces labor as per the elasticity of substitution. However, other research has argued that automation is labor-augmenting, where automation enhances the productivity of labor instead of capital (Bessen, 2018). Acemoglu and Restrepo (2018) reconcile the two opposing views by proposing a task-based framework where automation substitutes tasks previously performed by humans, but also can act as complements to labor as humans can then focus on more complex and creative subsequent tasks. In other words, they argue that machines cannot automatically be considered a replacement for human labor (substitutes) or facilitators for increased human productivity (complements); the impact of machine automation on human labor is specific to the task being considered.

We investigate whether machines can help augment human labor first, for the same and second, for sequential tasks. Empirically, if machines and humans are complements, we expect to find the largest reduction of risk premia to occur when human coverage and machine coverage are both high. On the other hand, if machines and humans are substitutes, we expect that the largest reduction of risk premia would occur in opposite corners i.e. when human coverage is low and machine coverage is high and vice-versa. Our portfolio tests show that the impact of machines is more pronounced in the absence of the human information retrieval activity in the EDGAR

database. We interpret this finding as machines acting as substitutes for human labor when performing the same task.

However, our tests on the role of machines in the presence of human agents who are tasked with more complex analysis (proxied by high institutional ownership and high stock analyst following) suggests that machines can also be viewed as complements to human effort. Specifically, in portfolio tests that adjust for firm size, we find the impact of machines is more pronounced for stocks that have high institutional ownership or stock analyst following. Taken together, our results paint a nuanced view on the interaction of machine automation and human labor (substitutes for the same task but complements in more complex environments) and is consistent with insights presented in Acemoglu and Restrepo (2018).

The rest of the paper proceeds as follows. Section 2 describes the data and construction of the machine coverage measure, section 3 discusses evidence relating machine coverage and stock returns, section 4 studies the interplay between human coverage and machine coverage, section 5 discusses possible broader interpretations of the machine coverage proxy, and section 6 concludes.

2. Data

Our main sample includes all common stocks on the NYSE, NASDAQ or AMEX stock exchanges from 2004 and 2016 with non-missing positive book equity, and firm data used in the summary statistics (Table 2) available on CRSP and Compustat. Further, in order to ensure that small stock illiquidity or bid-ask bounce are not driving our test results, we drop stocks with a price below $5 U.S. dollars, or stocks with a market capitalization in NYSE's bottom decile in the month prior to analysis.

Our main independent variable is sourced from the EDGAR Log File Data Set ("EDGAR Log"), recently made available to researchers by the SEC.[5] The database contains web server logs

[5] EDGAR Log File Dataset: https://www.sec.gov/dera/data/edgar-log-file-data-set.html

of internet search traffic for EDGAR filings on SEC.gov and is available from February 14, 2003. The log file records the timestamp (when), IP-address (who), file accessed (what), along with other information. The log data contains all access download requests, and as such contains requests from both humans and machines (i.e., computer programs or "robots").

Since our study is based on the relationship between machine automation in information processing and stock returns, an important first step is to separate machine downloads from human downloads. Table 1 Panel A provides an example of a suspected machine download in the EDGAR log. Table 1 Panel A1 lists the download activities of the IP=162.216.180.jde.[6] Within 3/21/2016 4:32:35, the IP-address accessed eight files in one second. We searched multiple geolocation databases and found that the IP-address is very likely to be located at Maverick Capital which is a long/short equity hedge fund in Dallas, Texas (Panel A2). Panel A3 shows the number of files downloaded over each minute on 3/21/2016. The IP has a very regular download pattern from 4 AM until midnight. In summary, Panel A shows that IP=162.216.180.jde is very unlikely to be a human on 3/21/2016.

2.1 Measuring machine coverage

Ryans (2017) investigates and compares various methods for parsing human and machine downloads in the EDGAR log. He proposes a method that resolves the disagreement among previous methods to provide an effective and cohesive classification of human and machine downloads. The proposed method is based on the author's detailed analysis of activity patterns by individual IP-addresses in the detailed log files and reduces both Type I and II errors in classification. Specifically, the Ryans (2017) filtering procedure to identify machines is based on the following three assumptions:

[6] The SEC masks the last three numbers of the IP-address with letters for anonymity (i.e., .jde in this case).

1. Humans do not download more than 25 items in a single minute;

2. Humans do not download more than 3 different companies' items in a single minute;

3. Humans do not download more than 500 items in a single day.

We use this classification method for our baseline result.[7] Specifically, we classify an IP address as human if the IP satisfies all the three conditions above. If it does not meet any of the above conditions, we define the IP as a machine for that day. We report basic summary statistics for machine and human downloads in the SEC database in Panel B of Table 1. For each quarter, the table reports the average daily values of the (1) total machine downloads, (2) total human downloads, (3) downloads per machine IP address, (4) downloads per human IP address, (5) machine downloads per firm, and (6) human downloads per firm. The average daily machine download increases 35-fold from 2004 Q1 to 2016 Q4 from 0.23 million to 8 million respectively.

Over the same period, the average daily human download changes from 0.11 million to 0.36 million, only a 3-fold increase. The moderate growth in human download is consistent with the intuition that human labor is scarce and limited, especially in the information and financial market where the skillset required is specialized. Importantly, we observe that machine downloads have increased to represent over 96% of total download requests as of 2016 Q4 (Figure 1 A). At the same time, we also see a large-scale increase in the share of equity market volume attributed to algorithmic or machine trading (Figure 1 B) from 20% in 2004 to 85% in 2012 (Glantz and Kissel, 2013). The rapid growth and dominant share of machine automation in accessing fundamental information has coincided with the growth in algorithmic or machine-attributed trading. While the time series, by itself, does not allow us to claim that automation in information collection results in

[7] Our test results are also robust to the earlier methods of machine-human classification used in the literature. Specifically, Loughran and McDonald (2016) separate machine and human IP addresses with the assumption that "Humans do not download more than 50 items during a day," while Drake, Roulstone, and Thornock (2015) assume that humans do not download more than 1,000 items during a day, and do not download more than 5 items per minute.

machine trading, the trend in both time series indicates that machine activity has come to increasingly dominate both information collection and trading activity over the past decade.

Our use of this machine coverage proxy to investigate the relation between machine automation and the financial market stems from four main reasons. First, while proprietary trading desks, hedge funds, etc. are more likely to use machines, they are also much more likely to be secretive and not reveal their in-house operations. Thus, the very fact that we can measure machine coverage using the EDGAR log provides a rare opportunity to empirically analyze its properties. Second, coverage is a revealed action-based measure that more precisely characterizes the nature of the information set covered by humans and machines, than a measure based on self-attestation of investor type (Abis, 2017; Birru, Gokkaya, Liu, 2018). Third, as classification is done on a daily basis our coverage measure is dynamic and considers the fact that an individual IP may be a human today but a machine tomorrow. Lastly, but most importantly, the presence of both machine and human agents in the EDGAR universe offers us a unique controlled environment to study the role of machines versus humans.

2.2 Determinants of machine coverage

We first calculate the daily number of machine IPs for each firm. We then aggregate these counts at a monthly level by taking the logarithm of the *average* number of daily machine IPs which download the firm's filings for the month (if there are no downloading machine IPs in a day, then the machine coverage for the day is 0). We next investigate the determinants of raw machine coverage in the cross-section by following Fang and Peress (2009), Hong, Lim, and Stein (2000), and Hillert, Jacobs, and Müller (2014) using a Fama-MacBeth regression as given below.

$$
\begin{aligned}
&Unajusted\ Machine\ Coverage_{s,t} \\
&= \beta_{1,t} Log\left(Size_{s,t}\right) + \beta_{2,t} SP500_{s,t} + \beta_{3,t} NASDAQ_{s,t} \\
&+ \beta_{4,t} Log\left(1 + \#Filing_{s,t}\right) + \beta_{5,t} Log\left(1 + \#Chars\ in\ Filing_{s,t}\right) \\
&+ \beta_{6,t} Log\left(1 + \#Analyst_{s,t}\right) + \beta_{7,t} Log\left(1 + \#News_{s,t}\right) + \beta_{0,t} \\
&+ \epsilon_{s,t} \qquad (1)
\end{aligned}
$$

Here, $Log\left(Size_{s,t}\right)$ is the market capitalization, $SP500_{s,t}$ is the indicator of the S&P 500 membership, $NASDAQ_{s,t}$ is the indicator of NASDAQ membership, and $Log\left(1 + \#Analyst_{s,t}\right)$ stands for the extent of analyst coverage. In addition to the explanatory variables in Hillert, Jacobs, and Müller (2014) above, we also include the degree of media coverage ($Log\left(1 + \#News_{s,t}\right)$) because news events might trigger information requests from the SEC. We also add two disclosure specific variables: the count of SEC filings, and the total size of SEC filings. The construction of the explanatory variables follows their respective standard literature and is described in detail in Appendix I.

We test the determinants of raw machine coverage in Table 3. Model 1 shows that firm size is an important determinant of raw machine coverage as it explains 10.6% of the cross-sectional variation by itself. Model 2 adds S&P 500 membership and NASDAQ membership as additional explanatory variables. Together they explain less than 0.2% of the additional variation of the machine coverage. Model 3 further includes the number of the SEC filings and the size of the filing as explanatory variables. These filing related variables lead to a large 15% increase in R-square of the model. Finally, in Model 4 we add media news and analyst presence which explains an additional 1% variation of in machine downloads.

Table 3 suggests the need to control for these machine coverage determinants such as firm size and information arrival (e.g. SEC filing and media news arrival) to isolate the intrinsic impact of the machine automation on the financial market. We therefore face the same challenge as Hong,

Lim and Stein (2000) who find that analyst coverage is directly correlated to firm size. We closely follow the approach of Hong, Lim, and Stein (2000), and Hillert, Jacobs, and Müller (2014) to isolate the net effect of our variable of interest – machine coverage. Specifically, we compute the residual machine coverage using equation (1) (as shown in Table 3 Model 4) as our proxy for *adjusted* machine coverage ("machine coverage" for short in the subsequent sections of the paper).[8]

3. Empirical analysis

3.1 Machine coverage and stock returns: Portfolio tests

In this section we test the cross-sectional relation between machine coverage and stock returns, the primary focus of our study. This examination allows us to answer whether machine activity is associated with increased or decreased risk premia in financial markets. A recent theoretical model by Begenau, Farboodi, and Veldkamp (2018) argues that big data and machines can help lower the equity premium. They argue that transformations in information technology have had a direct impact on financial markets. Specifically, the authors establish that both the availability of large amounts of data combined with the ability to process it allows market participants to better forecast firm value resulting in a reduction in risk for equity investment. In other words, machines act as positive agents as better information processing lowers the conditional variance and covariance of the firm's payoffs increasing their predictability, reducing their uncertainty, and thus lowering the cost of capital for the firm. Specifically, machines achieve this effect by making information available in a larger scale and quicker manner, and by additionally

[8] In Online Appendix 1, we document the persistence of adjusted machine coverage at the firm level for 1, 2, 3, 6, and 12-month intervals.

offering more efficient analysis. If this "*risk reduction*" hypothesis is true, we should see stocks with higher machine coverage exhibit lower returns as a result of lowered risk premia.

However, investors that automate trading activities are more likely to automate information collection as well. Recent estimates suggest that automated trading by highly sophisticated quantitative hedge funds accounts for approximately 29% of total trades in the United States.[9] A second opposing view sees the rise of machine-driven information processing through a more cautious lens. This is because unknown "*tail risks*" in machine automation can bring about unintended consequences and increase the risk of financial instability as seen in the flash crash of 2010 (Kirilenko and Lo, 2013; Kirilenko, Kyle, Samadi, and Tuzun, 2017). These left tail events can act to increase risk premia (Kelly and Jiang, 2014). Thus, if automation in information collection (machine coverage) is correlated with machine trading, which in turn exacerbates the risk premium, machine coverage can be associated with an increase in the cost of capital.

A third "*immateriality*" view posits that human attention and machine attention are no different, and thus should have similar end effects. Under this view machine coverage merely reflects the will of human analysts behind the computer. In this context, Veldkamp (2006) and Da, Engelberg, and Gao (2011) argue that human attention increases the trading demand for both buyer and sellers of an asset, while the selling demand is truncated by the short selling constraint (as it is always easier to buy than to sell short). The net effect is that asset prices can be inflated due to high human attention for a short period, before reversing course. In our context, this third view implies that machine coverage should have a similar effect to human coverage in the cross section of stock returns as machine attention is simply a mirror of human attention.

[9] https://www.straitstimes.com/world/united-states/automated-trading-and-wall-street-volatility

In summary, if the first "risk reduction" hypothesis is true, we should see stocks with higher machine coverage exhibit lower returns as a result of lowered risk premia. The second "tail risk" hypothesis has the exact opposite implication compared to the first hypothesis: stocks with higher machine coverage should exhibit higher returns due to increased risk premia. Like the second hypothesis, the third "immateriality" hypothesis also implies that higher machine coverage should exhibit higher returns, however unlike the second hypothesis where the price run-up should be permanent, the third hypothesis implies that the price run-up caused by attention should reverse due to subsequent correction of the mispricing driven by the behavioral bias.

Our study uses machine coverage as a proxy for the intensity of machine automation to shed new light on this critical debate. Machine coverage is the residual term of the monthly model given in equation (1) and described earlier in section 2.2. Each month, we sort firms into five quintiles of machine coverage and compute equally weighted returns following Fang and Peress (2009). The average monthly raw return in the lowest machine coverage quintile is 117 basis points, while that in the highest machine coverage quintile is 67 basis points. The difference in raw returns between the lowest and highest machine coverage quintile is economically and statistically significant at approximately 50 basis points per month, or about 6% annualized. Figure 2 plots the cumulative raw returns of portfolios by equally weighting stocks in the low machine coverage (long leg) and high machine coverage (short leg) quintiles from 2004 to 2016. Panel A documents the long and short legs separately, while Panel B plots the cumulative returns for the long minus short portfolio.

However, do the differences in returns between low and high machine coverage groups remain after adjusting for risk? We present results from Daniel, Grinblatt, Titman, and Wermers's (1997) [DGTW] characteristic based benchmarking procedure that allows firms to be precisely and reliably risk-adjusted for size, book-to-market and momentum characteristics than standard regression tests. DGTW adjusted returns in the second row of Table 4 present a similar pattern.

Overall, the average return premium between low coverage and high coverage stocks is almost unchanged at 51 basis points per month, or 6.12% per year. Figure 3 (structured like Figure 2 for raw returns) presents the cumulative DGTW adjusted returns over time. The Figure documents the consistent outperformance of the low machine coverage portfolio, even after the DGTW characteristic based risk adjustment, over time.

Next, we study alphas using standard factor adjustments across machine coverage quintiles. Row 3 exhibits alphas using the CAPM, row 4 using the 3-factor Fama-French (1993) model, row 5 using the updated Fama-French (2015) model that adds profitability and investment patterns, and row 6 using a 6-factor model that adds momentum to the 5-factor model presented in row 5. Our results are robust to these varying and rigorous factor specifications. We discuss 6-factor alphas below as that model poses the highest bar but results in other panels are similarly robust. Overall, the average difference in alphas between low coverage and high coverage stocks is approximately 58 basis points per month, or 6.96% per year.[10]

Taken together, the negative correlation between machine coverage and stock returns is robust to a wide variety of contemporary multi-factor risk adjustment methodologies, with magnitudes ranging from 6% to 7% annually across various tests for our sample period. These results are supportive of the first "risk reduction" hypothesis as stocks with lower machine coverage exhibit higher returns. Thus, our findings support the theoretical argument in Begenau, Farboodi, and Veldkamp (2018) that the availability of big data amplified by the increase in computing power can result in the reduction of the risk premium. At the same time, the strong negative correlation that we observe between machine coverage and subsequent returns does not support the "tail risk" and "immateriality" hypotheses.

[10] In Online Appendix 2, we show that our results are robust to portfolio formation periods of 1, 3, and 6 months in duration for the range of holding periods from 1-month to 6-months. Taken together, the test confirms that our results are not driven by short term reversals and is stable in the cross-section.

Firm size is an important determinant of the information environment surrounding a firm. Table 5 presents average monthly outperformance of stocks with low machine coverage against high machine coverage within each size group. Specifically, we present returns in low machine coverage quintile minus the high machine coverage quintile across five quintiles of firm size for each return measure studied in Table 4. We find that the difference between low and high machine coverage is significant for all except the biggest size quintile. Interestingly, this pattern across size groupings is similar to the one in Fang and Peress (2009)'s media news study. Their argument related to the significance of smaller stocks is particularly relevant in our case as well. They state that smaller stocks have poorer information dissemination compared to larger stocks that already have multiple information channels with sufficient manpower.

That said, we note that small stocks in our study are not particularly small, and our results are unlikely to be driven by the preponderance of very small firms. In particular, our filtering procedure eliminates stocks with prices below $5 at the beginning of the month, and stocks in the lowest decile of size following NYSE cut points. Furthermore, as noted in Table 5 firm size is controlled within a small range in each size quintile, while the level of machine coverage exhibits a wide variation. Thus, as the result we document is also present *in each of* the first four size quintiles, it is unlikely to be the size effect in disguise which would have been eliminated after controlling for size.

3.2 Machine coverage and stock returns: Fama-Macbeth regressions

In this section, we test the relation between machine coverage and stock returns in a robust multi-variate setting using Fama-Macbeth regressions. In particular, we model returns following the equation below where 'returns' are defined alternately a DGTW (1997) adjusted returns, CAPM alphas, Fama-French (1993) 3-factor alphas, Fama-French (2015) 5-factor alphas, and finally alphas from a model combining the Fama-French (2015) 5-factors and the Carhart (1997) momentum factor.

$$Ret_{s,t} = \beta_{machine} Machine\ Coverage_{s,t-1} + \beta_1 r_{s,t-1} + \beta_2 r_{s,t-2,t-12} + \beta_3 Volume_{s,t-1} + \beta_4 Log(Size_{s,t}) + \beta_5 SP500_{s,t} + \beta_6 NASDAQ_{s,t} + \beta_7 Log(1 + \#Analyst_{s,t}) + \beta_8 Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t} \quad (2)$$

The Fama-Macbeth setup allows us to control for a variety of autocorrelation terms (return in month t-1, return from t-12 to t-2) and firm characteristics (size, monthly trading volume, S&P and NASDAQ dummy variables, and the extent of analyst and news coverage). The coefficients, presented in Table 6, represent time series averages of monthly model estimates for the sample period from 2003 to 2016. Model (5) indicates that a one standard deviation increase in machine coverage results in a -1.08% [-12.96% annualized] change in 6-factor alphas in the following month. Similar values are observed in DGTW, CAPM, 3-factor, and 5-factor tests presented in Models (1) through (4). The Fama-Macbeth results corroborate our earlier findings in the portfolio tests, as the coefficient on machine coverage is negative and highly significant across the various definitions of returns both in the presence and absence of various risk-adjustments.

3.3 XBRL adoption and machine coverage: A natural experiment

The SEC mandated that financial statements be simultaneously filed using XBRL (eXtensible Business Reporting Language) in 2009 to improve the transparency of financial disclosure and reduce information processing costs (Bartley et al., 2011; Blankespoor, 2019). XBRL is a variant of XML (eXtensible Markup Language) and is designed to allow corporate disclosure to take place in a structured machine-readable format. The adoption of XBRL was intended to facilitate easier search and retrieval as disclosure filings and various accounting data items within them would come pre-tagged for machine processing.

The XBRL mandate was implemented in 3 phases. The first set of firms began XBRL disclosure for the period ending on or after June 15, 2009, a second and third set followed exactly

in one-year intervals at June 15, 2010 and June 15, 2011. We use the staggered implementation of the XBRL regime as a natural external shock following Dong et al. (2016) to investigate the causal association between machine coverage and stock returns. In our context, the implementation of the XBRL coding enables machines to quickly and precisely parse information for analysis, potentially enhancing their utility.

Given the multiple-stage staggered nature of the external shock, we follow the modified difference-in-difference setup of Gao and Huang (2018) who utilize the multiple-stage phase-in schedule for the implementation of the EDGAR system to study the role of information dissemination technologies.

$$\begin{aligned} Ret_{s,t} = {} & \boldsymbol{\beta_0} + \boldsymbol{\beta_{XBRL\times machine}}\boldsymbol{Post\ XBRL_{s,t}} \times \boldsymbol{Machine\ Coverage_{s,t-1}} \\ & + \boldsymbol{\beta_{machine}}\boldsymbol{Machine\ Coverage_{s,t-1}} + \boldsymbol{\beta_{XBRL}}\boldsymbol{Post\ XBRL_{s,t}} + \beta_1 r_{s,t-1} \\ & + \beta_2 r_{s,t-2,t-12} + \beta_3 Volume_{s,t-1} + \beta_4 Log\left(Size_{s,t-1}\right) + \beta_5 NASDAQ_{s,t-1} \\ & + \beta_6 SP500_{s,t-1} + \beta_7 Log\left(1 + Analyst_{s,t-1}\right) + \beta_8 Log\left(1 + News_{s,t-1}\right) + \epsilon_{r,t} \end{aligned}$$

Specifically, we create a dummy variable called Post XBRL which is set to 1 when a firm first starts reporting using XBRL, and is equal to 0 otherwise. We interact this Post XBRL dummy with our machine coverage measure in the prior month and study the impact of this interaction on DGTW, CAPM, 3-factor, 5-factor, and 6-factor adjusted return. Our previous tests have shown that machine coverage has a negative relation with the cross section of stock returns. If machines are involved in parsing text, we hypothesize that the interaction term between machine coverage and Post XBRL indicator would be negative and significant since XBRL would enhance the ability of machines to parse information. The regression design includes fixed effects at the firm and time (monthly) level to control for time-invariant differences across firms and aggregate fluctuations in stock returns. In line with our Fama-Macbeth tests in Table 6, we continue to control for short term (last month) and long term (prior year) returns, volume, firm size, NASDAQ membership, S&P 500 membership, and the extent of analyst and news coverage. However, as this test involves a

pooled panel regression, we additionally cluster standard errors by firm and time following Peterson (2009).

Table 7 reports the results of this difference-in-difference test. The base machine coverage measure remains negative and highly significant. Importantly, in regard to the difference-in-difference test, we find that the interaction term (Post XBRL*Machine Coverage) is negative and significant for all the various factor adjusted return measures we study from DGTW returns to 6-factor adjusted returns.

Specifically, the coefficient on machine coverage by itself is -28 basis points, while the interaction of machine coverage and the Post XBRL dummy is -45 basis points in the 6-factor specification. In other words, the impact of machine coverage on returns increased from -0.28% prior to the introduction of XBRL to -0.73% in the post-XBRL regime. Both variables are significant at the 99% confidence level. This implies that after the introduction of the XBRL regime the impact of machine coverage on stock returns has more than doubled. The result shows that an increase in the ability of machines to parse information (instrumented by the implementation of the XBRL regime) is associated with a strengthening of the machine and stock return relationship, implying a causal relationship between machine coverage and stock returns. Finally, results from the difference-in-difference test also allow us alleviate concerns related to endogeneity and potentially spurious results due to the presence of firm-specific time-invariant omitted variables.

4. What is the role of human analysis in the presence of machines?

The above results show that non-human activities, specifically machine automation in information processing, directly impacts the financial market with effects visible in the cross-section of stock returns. The findings are important given the increasing share of information being analyzed by machines. Our results indicate that at least in the financial marketplace, machine automation is already significantly associated with information discovery and thus the return

generating process. Indeed, at BlackRock (the world's largest asset management firm) human managers are being replaced with machines that rely on algorithms and models to pick stocks (Thomas, 2017).[11] We now ask what will be the role of human analysis in this new reality, and will the presence of machines alter the scope and character of human work?

Recent theoretical research in economics has proposed various frameworks to model automation. Some have modeled automation as a factor-augmenting force, factors being the standard ones in the Cobb-Douglas production function (i.e., capital and labor). For instance, Sachs and Kotlikoff (2012), Graetz and Michaels (2015), and Nordhaus (2015) work under the premise that automation is capital-augmenting as it increases the productivity and thus the return to capital. Yet others have argued that automation is labor-augmenting, which results in increased productivity of labor. In this view, automation increases labor demand and wages and argues against the assertation that human employment is at risk with the rise of new information technologies (Bessen, 2018). These two views are disparate, with the former hypothesis seeing automation as a substitute for labor, while the latter seeing it as a complementary force.

Acemoglu and Restrepo (2018) reconcile the two views by presenting a task-based view on the interaction of machine automation and human labor. Their framework suggests while tasks previously performed by labor will be automated, machines can complement humans to perform new tasks in more complex settings where humans have a comparative advantage. In other words, they argue that machines cannot automatically be considered as substitutes for human labor or complements that increase human productivity. The impact of machine automation on human labor is specific to the task being considered.[12]

---

[11] Thomas, Landon Jr. (New York Times, 2017)
https://www.nytimes.com/2017/03/28/business/dealbook/blackrock-actively-managed-funds-computer-models.html

[12] In this popular press, Daugherty and Wilson (2018) present a symbiotic model of the way robots and artificial intelligence will augment human workers in what they term *human+machine* collaboration. In their

We first test whether humans and machines are complements or substitutes for the same information processing task related to processing EDGAR documents. Importantly, our tests are conducted in a controlled financial market setting where both humans and machines co-exist. First, our setting provides the same information environment and resource (i.e. EDGAR file retrieval system) for both humans and machines to work on. Second, the specific task undertaken by humans and machines is similar (i.e. to retrieve information). And third, there is a common measure of the outcome (i.e. risk premium reduction) to judge the contribution of both humans and machines. The degree of machine automation in our context can be modeled using the extent of machine coverage, while labor can be proxied for by human coverage in EDGAR. We define (adjusted) human coverage analogously to (adjusted) machine coverage.[13]

We additionally investigate whether machines can help augment human labor for sequential or different tasks in section 4.2. We choose two proxies for the presence of human labor who may have an advantage in dealing with complexity compared to machines, and where humans may have additional tasks such as analysis and trading beyond basic information collection. In doing so, we can test whether Acemoglu and Restrepo (2018)'s task based nuanced view on the interplay of machines and humans (complements in some tasks, substitutes in others) is supported in our data.

4.1 Relation between humans and machines for the same information processing task

view, machines will not make humans redundant in the workplace, but the nature of human work has to be re-imagined as humans and machines collaborate more closely.

[13] Adjusted human coverage ($Human_{s,t}$) of stock *s* in month t is estimated as the residual term of the monthly model: $Human\ Coverage_{s,t} = \beta_{1,t} Log(Size_{s,t}) + \beta_{2,t} SP500_{s,t} + \beta_{3,t} NASDAQ + \beta_{4,t} Log(1 + \#Filing_{s,t}) + \beta_{5,t} Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t} Log(1 + \#Analyst_{s,t}) + \beta_{7,t} Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$, where Human Coverage (Unadjusted) of month t is estimated as the logarithm of the average number of daily human IPs which download filings of firm s over calendar days in month t (if there is no downloading machine IPs in a day, then the human coverage of the day is 0). Following Ryan (2017), we label an IP as a machine IP if any of the following condition is met: 1. The daily download is more than 500. 2. The per-minute download is more than 25. 3. The per-minute firms download is more than 3. The human IPs are defined as the non-machine IPs.

In our primary test to examine this issue we compute 6-factor alphas where factors include the Fama-French (2015) 5-factors plus the Carhart (1997) momentum factor in a 3x5 double sort of human coverage (terciles) and machine coverage (quintiles). If humans and machines are complements as argued by the labor-augmenting models, we expect the largest reduction in risk premium to occur when human and machine coverage are both high. If on the other hand, humans and machines are substitutes as described in the capital-augmenting models, we expect that the largest reduction of risk premium to occur in opposite corners. In other words, the difference between the low and high machine coverage quintiles should be greatest when human coverage is low, and the difference between low and high human coverage terciles should be greatest when machine coverage is low.

Our findings, presented in Table 9 Panel A, indicate that stocks with low machine coverage outperform those with high machine coverage only in the low human coverage tercile. If human activity is similar to machine activity, then high human coverage should mitigate the impact of machine coverage and we should not observe a premium reduction attributed to machines. This finding is consistent with the notion that human and machines are performing similar tasks and are *substitutes* in this context. In the opposite scenario, we find that stocks with low human coverage outperform those with high human coverage in the low machine coverage quintile, and not in the high machine coverage quintile. The difference between low and high human coverage stocks in the low machine coverage quintile is once again large and highly significant at 11.8% annualized. Thus, in both opposing corners we find evidence supporting the hypothesis that machines are substitutes for human labor in this task of information retrieval.

As an analogue to the portfolio tests, we repeat the Fama-Macbeth experiment from section 3.2 but include both human and machine coverage as separate terms in Table 9 Panel B. We find that both human and machine coverage are negative and significant in the presence of each other. The coefficients indicate that a one standard deviation increase in machine (human) coverage

results in a -0.58% (-0.36%) [-8.56% (-5.79%) annualized] change in 6-factor alphas in the following month. The interaction term is positive and significant across all specifications which suggests that human coverage and machine coverage are substitutes for the same information processing task.

4.2 Relation between humans and machines for different information processing tasks

We now examine the role of machines in the presence of human agents such as institutional investors and stock analysts who are likely to perform cognitively more complex and creative tasks. In portfolio tests that adjust for firm size, we find that the impact of machines is more pronounced in settings with high institutional ownership or high stock analyst following. This suggests that machines also act as complements to tasks performed by human agents. Specifically, in Table 10 we present results from a triple sort on firm size, machine coverage (quintiles) and institutional ownership in Panel A and firm, size, machine coverage, and stock analyst coverage in Panel B. Panel A shows that the difference between the low machine coverage and high machine coverage portfolio return approximately doubles from 41 basis points per month in the low institutional ownership tercile to 81 basis points in the high institutional ownership tercile. In similar vein, Panel B shows that the difference between the low machine coverage and high machine coverage portfolio return goes from 35 basis points per month in the low analyst coverage tercile to 79 basis points per month in the middle analyst coverage tercile to 193 basis points per month in the top analyst coverage tercile.

Taken together, our results paint a more realistic picture of the current role of machine automation in financial markets as both substitutes and complements depending on the specific tasks; as we find that machines and human labor are substitutes for the same information processing task (downloads), while being complements in settings with sequential and more complex information processing needs as shown above.

5. Discussion on the interpretation of the "Machine Coverage" proxy

Automation in information processing can span a range of activities from information collection, storage, shallow analysis (e.g. data cleansing, entity tagging, text parsing, summary statistics, etc.) to deeper analysis (e.g. prediction, classification, etc.) leading up to automated trading. Our machine coverage proxy is based on machine downloads, which maps most directly into the early stage of the information process related to information collection and storage. For the current discussion, we view this as the narrow interpretation of the machine coverage proxy. However, machine coverage might also proxy for the later stage in information processing related to shallow and deep analytics as agents that automate collection are also likely to automate analysis. We view this as the broader interpretation of the machine coverage proxy.

In this section, we test to what extent our empirical machine coverage *proxy* is consistent with a broader view of machine automation in financial markets. We begin by noting that under the narrow view the machine's *job* ends at document retrieval and does not include document analysis (i.e. the machine does not "look" into the content of the downloaded document). In other words, if machine coverage only mapped into the narrow definition of information processing related to information collection, double sorts based on factors related to characteristics of the downloaded content, would not have any bearing on the cross-sectional effects that we observe across machine coverage groups. However, if we find that this is not the case, it is likely that a simple machine coverage measure based on downloads can serve as a good proxy for a broader interpretation of machine automation and information processing.

We begin by examining the readability of the financial disclosure as it directly relates to the content of the information obtained. We sort firm disclosures into five quintiles using a readability index based on eight commonly used readability measures (Li, 2008). These include the Flesch reading ease index, the Flesh-Kincaid grade level index, the RIX index, the Coleman-Liau index, the Gunning-Fox index, the ARI index, the SMOG index and the LIX index (Appendix 1). We then

apply a 5x3 double sort with machine coverage (x5) and readability (x3) to test our hypothesis. Results presented in Table 11 show that the difference in 6-factor alphas between low and high machine coverage is greater in high readability terciles, suggesting that machines add more value when the underlying texts are more readable. Specifically, the average annualized difference in 6-factor alphas between low and high machine coverage quintiles among high readability firms is 11.08%, while the difference among the low readability firms is 4.50%. These results show that the content of information has a direct bearing on the cross-sectional relation between machine coverage and stock returns.

In this context, we note that tests in section 3.3 on XBRL also shed light on the broad interpretation of the proxy. One could argue that the natural experiment on XBRL is a joint test that (1) machine coverage has a causal relation with stock returns, and (2) machines are involved with parsing the underlying text. The parsing of text is an element in the information processing chain under the broad view. This is because text parsing is a step beyond the downloading of data and can involve advanced natural language processing techniques which XBRL makes easier. If machine automation is not involved in text parsing the introduction of XBRL should lead to no effect on machine's coverage relation with the cross section of stock returns. However, since we find that the introduction of XBRL impacts the relation of machine coverage on the cross section of stock returns, we conclude that this evidence points to the broader view of the machine coverage proxy.

While each test individually is not conclusive, taken together the results are indicative of the broader interpretation of the machine coverage proxy for some tasks in the information chain.

6. Conclusion

Investors use a variety of sources in constructing their information set about firms in which they are interested. The Security and Exchange Commission's (SEC) Electronic Data Gathering

and Retrieval (EDGAR) database represents a potent primary source of new information as it collates, disseminates, and archives financial statements along with other regulatory disclosures for the full set of publicly listed firms in the United States. The increased magnitude and complexity of public firm disclosure over the years coupled with an increased availability of computing power has dramatically changed the way in which investors process public disclosure. We study the impact of a significant change – the domination of machines in primary information collection – on risk premia evidenced in the cross-section of stock returns.

We find that stocks in the lowest quintile of machine coverage outperform those in the highest quintile by 6 to 7% annually after adjusting for risk. Our results are consistent with recent theoretical work on big data by Begenau, Farboodi, and Veldkamp (2018) since higher machine coverage significantly reduces the risk premium. Second, we confirm that our findings are also present both in portfolio tests and Fama-Macbeth style regressions where returns are measured as DGTW returns, or as CAPM, 3-factor, 5-factor, and 6-factor alphas. Finally, we confirm our results in a natural experiment related to the phased introduction of XBRL machine readable financial disclosure that increased the ability of machines to process information.

Our findings are consistent with the theoretical work of Acemoglu and Restrepo (2018) as we find that machine and human labor are substitutes for the same information processing task, while being complements in settings with sequential or more complex information processing needs. Overall, our findings shed light on the rapidly evolving nature with which fundamental information is processed in financial markets today.

References

Abis, S., 2017. Man vs. machine: quantitative and discretionary equity management. *Unpublished working paper. Columbia Business School.*

Acemoglu, D., and Restrepo, P., 2018. Modeling automation. *National Bureau of Economic Research working paper no. w24321.*

Barber, Brad M., and Terrance Odean., 2007. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies* 21(2), pp. 785-818.

Bartley, J., Chen, A., Taylor, E., 2011, A comparison of XBRL filings to corporate 10-Ks – Evidence from the voluntary filing program, *Accounting Horizons 25(2),* pp. 227-245.

Begenau, J., Farboodi, M., and Veldkamp, L., 2018. Big data in finance and the growth of large firms. *Journal of Monetary Economics*.

Ben-Rephael, A., Da, Z., and Israelsen, R.D., 2017. It depends on where you search: institutional investor attention and underreaction to news. *The Review of Financial Studies, 30*(9), pp. 3009-3047.

Bessen, J.E., 2018. Automation and jobs: When technology boosts employment. *Boston University School of Law Research Paper No. 17-09.*

Blankespoor, E. 2019. The impact of information processing costs on firm disclosure choice: evidence from the XBRL mandate. *Journal of Accounting Research, 57(4),* pp. 919-967.

Birru, J., Gokkaya, S., and Liu, X., 2018. Capital market anomalies and quantitative research. *Fisher College of Business working paper no. 2018-03-007.*

Carhart, M.M., 1997. On persistence in mutual fund performance. *The Journal of Finance, 52*(1), pp.57-82.

Daniel, K., Grinblatt, M., Titman, S. and Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. *The Journal of Finance, 52*(3), pp.1035-1058.

Daugherty, P.R., and Wilson, H.J., 2018. Human + Machine: Reimagining work in the age of AI. *Harvard Business Press*.

Dong,Y., Li, O., Lin, Y., and Ni, C, 2016. Does information-processing cost affect firm-specific information acquisition? Evidence from XBRL adoption. *Journal of Financial and Quantitative Analysis 51(2),* pp. 435-462.

Drake, M.S., Roulstone, D.T., and Thornock, J.R., 2015. The determinants and consequences of information acquisition via EDGAR. *Contemporary Accounting Research, 32*(3), pp. 1128-1161.

Da, Z., Engelberg, J., and Gao, P., 2011. In search of attention. *The Journal of Finance, 66*(5), pp. 1461-1499.

Fama, E.F. and French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), pp.3-56.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), pp. 1–22.

Fang, L., and Peress, J., 2009. Media coverage and the cross-section of stock returns. *The Journal of Finance, 64*(5), pp. 2023-2052.

Frankel, R., Kothari, S.P., and Weber, J., 2006. Determinants of the informativeness of analyst research. *Journal of Accounting and Economics, 41*(1-2), pp. 29-54.

Gao, M. and Huang, J., 2018. Informing the market: The effect of modern information technologies on information production. *University of Illinois at Urbana-Champagne working paper.*

Glantz, M., and Kissel, R., 2013. Multi-asset risk modeling: Techniques for a global economy in an electronic and algorithmic trading era. *Academic Press,* pp. 258.

Graetz, G. and Michaels, G., 2015. Robots at work. *CEPR discussion paper no. DP10477.*

Hillert, A., Jacobs, H., and Müller, S., 2014. Media makes momentum. *Review of Financial Studies, 27*(12) pp. 3467–3501.

Hong, H.G., Lim, T., and Stein, J.C., 2000. Bad news travels slowly: size, analyst coverage and the profitability of momentum strategies. *Journal of Finance*, 55(1), pp. 265–95.

Kelly, B., and Jiang, H., 2014. Tail risk and asset prices. *The Review of Financial Studies, 27*(10), pp. 2841-2871.

Kirilenko, A., Kyle, A.S., Samadi, M., and Tuzun, T., 2017. The flash crash: high-frequency trading in an electronic market. *The Journal of Finance, 72*(3), pp. 967-998.

Kirilenko, A., and Lo, A.W., 2013. Moore's law versus murphy's law: algorithmic trading and its discontents. *Journal of Economic Perspectives,* 27(2), pp. 51-72.

Li, F., 2008. Annual Report Readability, Current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), pp. 221 – 247.

Loughran, T., and McDonald, B., 2017. The use of EDGAR filings by investors. *Journal of Behavioral Finance, 18*(2), pp. 231-248.

Nordhaus, W.D., 2015. Are we approaching an economic singularity? Information technology and the future of economic growth. *National Bureau of Economic Research working paper no. w21547.*

Peterson, M., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies 22,* pp. 435-480.

Ryans, J., 2017. Using the EDGAR log file data set. *London Business School working paper. http://www.jamesryans.com/*

Sachs, J.D. and Kotlikoff, L.J., 2012. Smart machines and long-term misery. *National Bureau of Economic Research working paper no. w18629.*

Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), pp. 1437-1467.

Veldkamp, L., 2006. Media frenzies in markets for financial information. *American Economic Review, 96*(3), pp. 577-601.

Appendix 1

Variable Descriptions

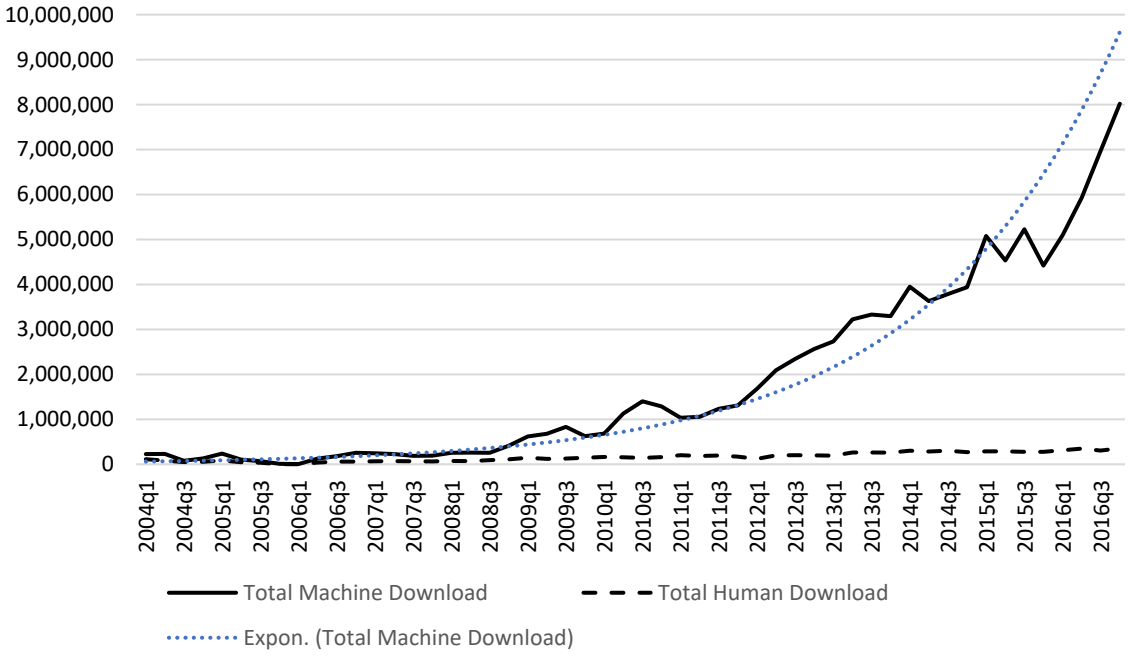
Variable	Description
Financial Market	
Log(Size)	Logarithm of the market capitalization.
S&P500	Indicator equal to 1 if the firm is ann S&P500 member and 0 otherwise.
NASDAQ	Indicator equal to 1 if the firm is listed on NASDAQ and 0 otherwise.
Book to Market Equity	Book equity to market equity ratio.
SEC Filing	
Log(1+#Filings)	Natural logarithm of the (1+number of filings). We consider all the filings in the month.
Log(1+#Characters in Filings)	Natural logarithm of the (1+total characters in all the filings). We consider all the filings in the month.
Readability	We estimate SEC sentiment as the total characters (in the filings) weighted (Flesch_Reading_Ease+Flesch_Kincaid_Grade_Level+RIX+Coleman_Liau_Index+Gunning_Fog_Index+ARI+SMOG_Index+LIX)/8, where the terms are the readability indices (Li 2008), and each of the indexes is standardized by subtracting its monthly mean and divided its monthly standard deviation
Analyst and Media	
Log(1+#Analyst)	Following Hong, Lim and Stein (2000), analyst coverage is the natural logarithm of (1+number of estimates for the firm's earnings next year). We consider the value in the given month.
Log(1+#News)	News coverage is the natural logarithm of (1+number of news articles). A news article is counted towards a firm if the relevance score of the article is above 90 (out of 100). We consider the value in the given month.
EDGAR Download	
Machine	Following Hillert, Jacobs, and Müller (2014)'s adjustment to the media coverage, the adjusted machine coverage ($Machine_{s,t}$) of stock s in month t is estimated as the residual term of the monthly model: $Machine\ Coverage_{s,t} = \beta_{1,t} Log(Size_{s,t}) + \beta_{2,t} SP500_{s,t} + \beta_{3,t} NASDAQ + \beta_{4,t} Log(1 + \#Filing_{s,t}) + \beta_{5,t} Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t} Log(1 + \#Analyst_{s,t}) + \beta_{7,t} Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$, where Machine Coverage (Unadjusted) of month t is estimated as the logarithm of the average number of daily machine IPs which download filings of firm s over calendar days in month t (if there is no downloading machine IPs in a day, then the machine coverage of the day is 0). Following Ryan (2017), we label an IP as a machine IP if any of the following condition is met: 1. The daily

	download is more than 500. 2. The per-minute download is more than 25. 3. The per-minute firms download is more than 3.
Human	Following Hillert, Jacobs, and Müller (2014)'s adjustment to the media coverage, the adjusted human coverage ($Human_{s,t}$) of stock s in month t is estimated as the residual term of the monthly model: $Human\ Coverage_{s,t} = \beta_{1,t}Log(Size_{s,t}) + \beta_{2,t}SP500_{s,t} + \beta_{3,t}NASDAQ + \beta_{4,t}Log(1 + \#Filing_{s,t}) + \beta_{5,t}Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t}Log(1 + \#Analyst_{s,t}) + \beta_{7,t}Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$, where Human Coverage (Unadjusted) of month t is estimated as the logarithm of the average number of daily human IPs which download filings of firm s over calendar days in month t (if there is no downloading machine IPs in a day, then the human coverage of the day is 0). Following Ryan (2017), we label an IP as a machine IP if any of the following condition is met: 1. The daily download is more than 500. 2. The per-minute download is more than 25. 3. The per-minute firms download is more than 3. The human IPs are defined as the non-machine IPs.

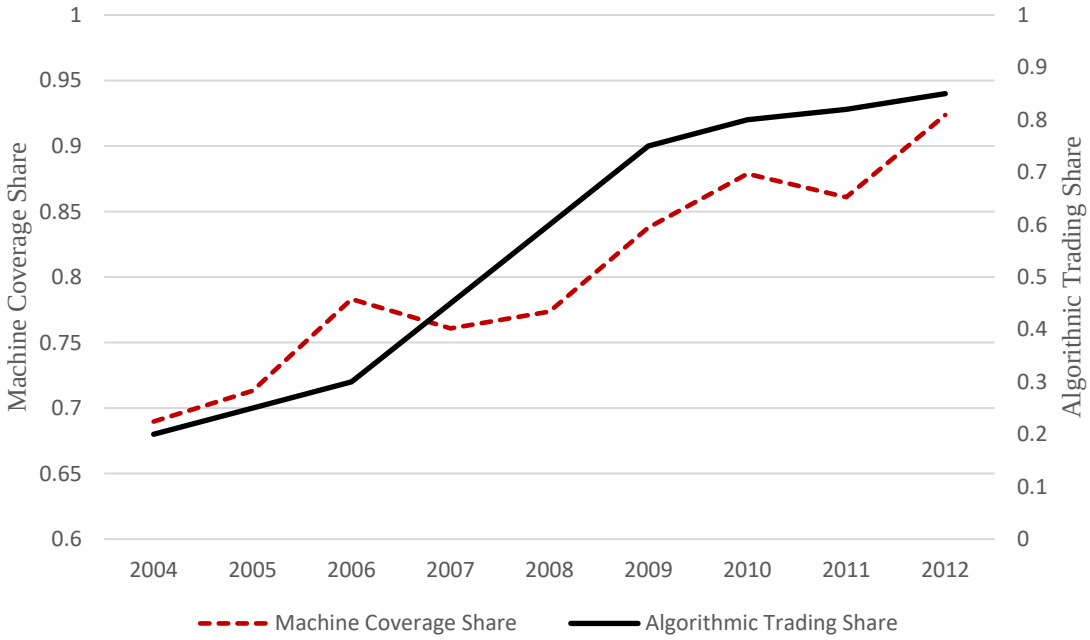
Figure 1

Aggregate Machine and Human Coverage Over Time

In Panel A, we plot the aggregate quarterly monthly machine and human downloads from the SEC EDGAR website from 2004 Q1 to 2016 Q4. The black line represents total machine downloads, while the dashed line represents total human downloads. The dotted line represents an exponential fitted line on total machine downloads. In Panel B, we plot the *share* of total downloads attributed to machines along with the percentage of total stock market volume attributed to algorithms. The data for machine coverage share is based on our analysis from Panel A, while estimates of algorithmic trading are from Glantz and Kissel (2013).



Panel A

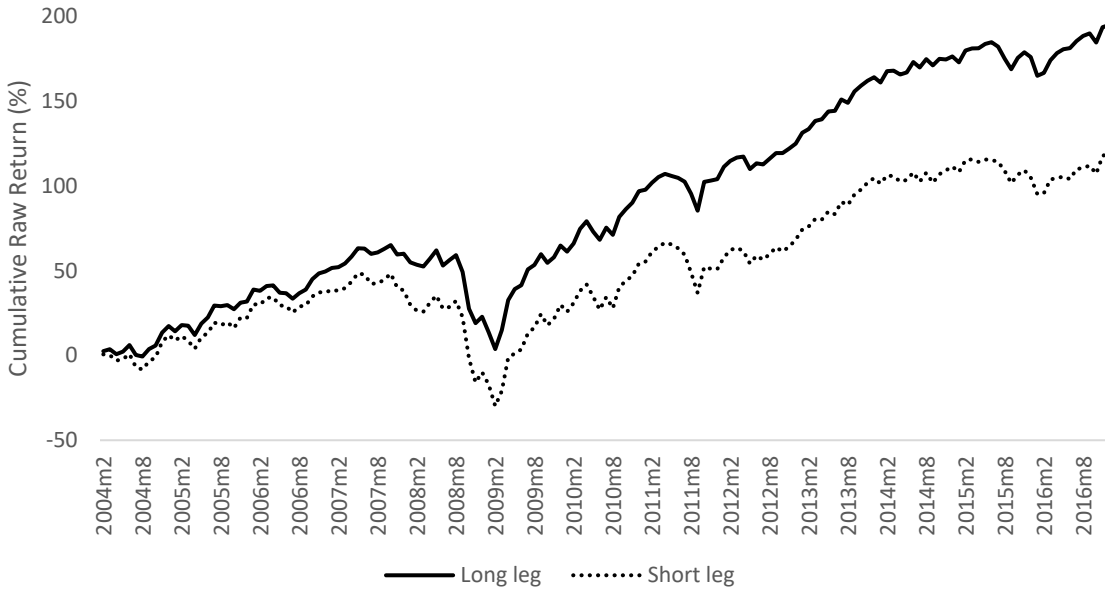


Panel B

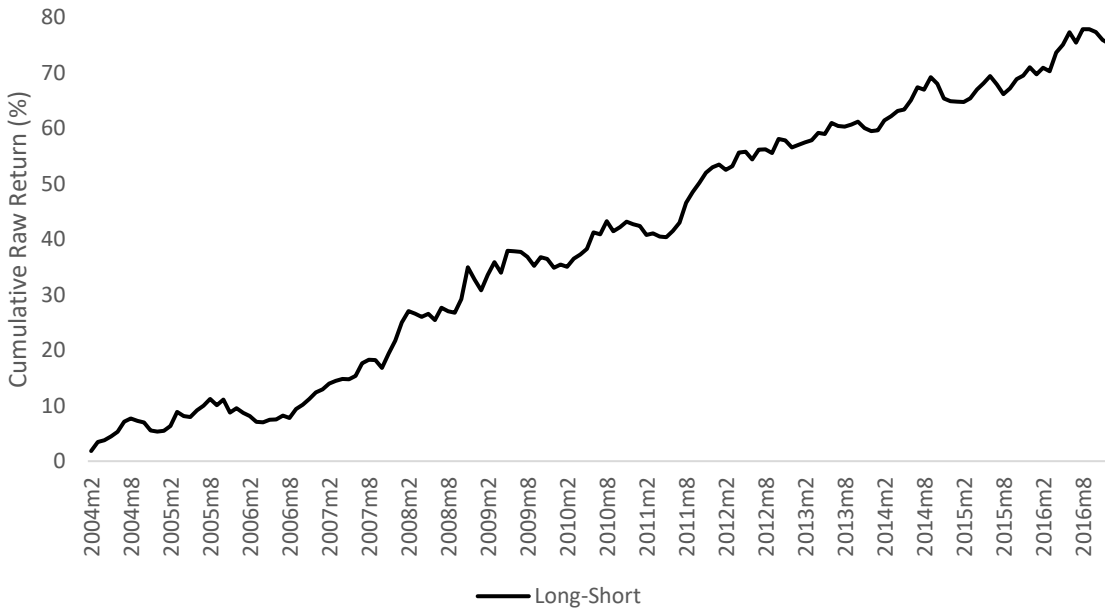
Figure 2

Cumulative Raw Returns

In this figure we plot cumulative raw returns (in percentages) of portfolios created by equally weighting stocks in the low machine coverage (long leg) and high machine coverage (short leg) quintiles from 2004 Q2 to 2016 Q4. Machine coverage is measured as of the prior month, and portfolios are rebalanced monthly. Panel A shows returns to the long and short portfolios separately, while panel B depicts returns to the zero cost long-short portfolio.



Panel A

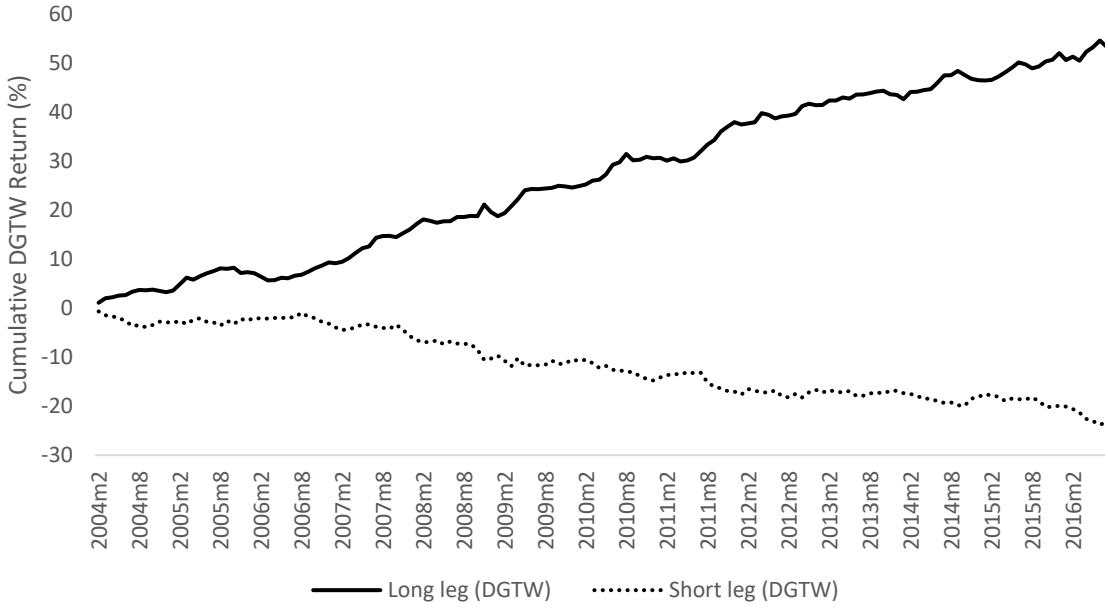


Panel B

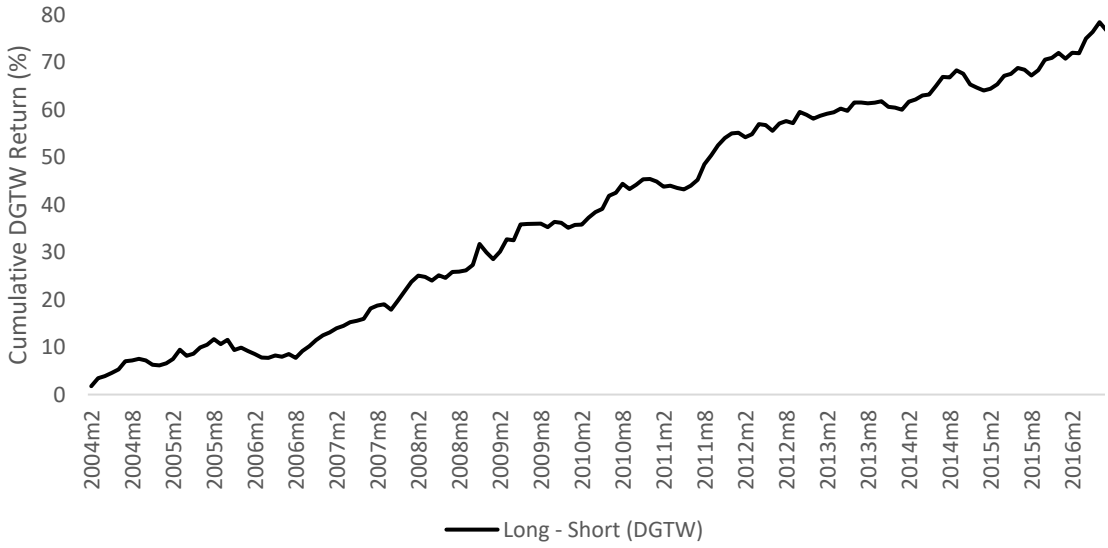
Figure 3

Cumulative DGTW Returns

In this figure we plot cumulative Daniel, Grinblatt, Titman, and Wermers (1997) [DGTW] adjusted returns (in percentages) of portfolios created by equally weighting stocks in the low machine coverage (long leg) and high machine coverage (short leg) quintiles from 2004 Q2 to 2016 Q4. The DGTW procedure is a characteristic based benchmarking procedure that adjusts size, book-to-market, and momentum. Machine coverage is measured as of the prior month, and portfolios are rebalanced monthly. Panel A shows DGTW returns to the long and short portfolios separately, while panel B depicts DGTW returns to the zero cost long-short portfolio.



Panel A



Panel B

Table 1: Summary of EDGAR Downloads

This table describes the EDGAR Web Log dataset. Panel A provides an example of an EDGAR download. Specifically, Panel A.1 reports a small segment of the downloads by IP address 162.216.180.jde (i.e., IP address with the last three digits masked). The data includes the IP address, the download date and time, the company identifier (CIK), file identifier (Accession) along with other information. Panel A.2 shows the geographical location and the possible identity of the IP address obtained from a reverse IP lookup database. The IP address of 162.216.180.jde is likely used by Maverick Capital, which is a long/short equity hedge fund in Dallas. Panel A.3 reports the EDGAR download volume by 162.216.180.jde on 2016-03-21. Panel B summarizes the basic statistics of the EDGAR download. Following Ryans (2017), we classify an IP as not a human in a given day if at least one of the following condition is met: 1. The daily download is more than 500. 2. The per-minute download is more than 25. 3. The per-minute number of firms downloaded is more than 3. The rest of the IPs are classified as human downloads. Specifically, Panel B reports the daily average of the total machine download, total human download, per machine download, per human download, per firm machine download, and per firm human download over each quarter. The sample period is 2004-2016.

Panel A.1: Example of the Downloads by IP=162.216.180.jde

IP	Date	Time	CIK	Accession	Size
162.216.180.jde	3/21/2016	4:32:34	706688	0000706688-16-000262	5766
162.216.180.jde	3/21/2016	4:32:34	1551138	0001638599-16-000749	72735
162.216.180.jde	3/21/2016	4:32:34	1551139	0001638599-16-000749	72735
162.216.180.jde	3/21/2016	4:32:34	1591588	0001591588-16-000143	7383
162.216.180.jde	3/21/2016	4:32:35	706688	0000706688-16-000264	5079
162.216.180.jde	3/21/2016	4:32:35	706688	0000706688-16-000263	5762
162.216.180.jde	3/21/2016	4:32:35	1018840	0001225208-16-030090	5801
162.216.180.jde	3/21/2016	4:32:35	706688	0000706688-16-000265	5067
162.216.180.jde	3/21/2016	4:32:35	706688	0000706688-16-000267	5722
162.216.180.jde	3/21/2016	4:32:35	706688	0000706688-16-000266	5084
162.216.180.jde	3/21/2016	4:32:35	1599103	0000899243-16-016223	5710
162.216.180.jde	3/21/2016	4:32:35	1408175	0001246360-16-005174	7404
162.216.180.jde	3/21/2016	4:32:36	1018840	0001225208-16-030091	5910
162.216.180.jde	3/21/2016	4:32:36	1018840	0001225208-16-030089	5679
162.216.180.jde	3/21/2016	4:32:36	1400956	0001400956-16-000031	5722

Panel A.2: Identity of IP=162.216.180.jde

Geolocation data from ipinfo.io (Product: API, real-time)

IP Address	Country	Region	City
162.216.180.100	United States	Texas	Dallas
ISP	**Organization**	**Latitude**	**Longitude**
Maverick Capital, LTD.	Maverick Capital, LTD.	32.7877	-96.8000

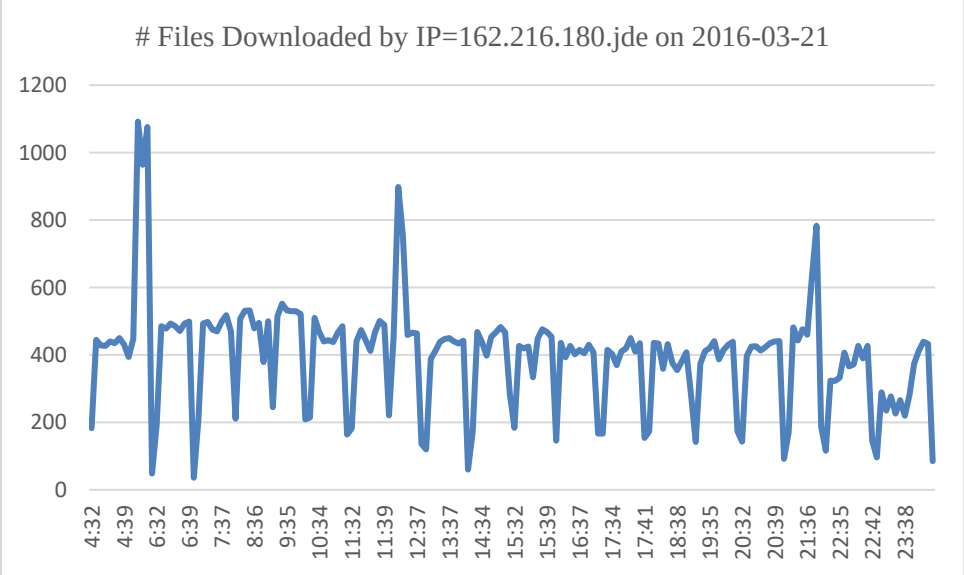
Geolocation data from EurekAPI (Product: API, real-time)

IP Address	Country	Region	City
162.216.180.100	United States	New York	Brooklyn
ISP	**Organization**	**Latitude**	**Longitude**
Maverick Capital	Maverick Capital	40.6944	-73.9906

Geolocation data from DB-IP (Product: Full, 2018-5-2)

IP Address	Country	Region	City
162.216.180.100	United States	Texas	Dallas (Uptown)
ISP	**Organization**	**Latitude**	**Longitude**
Maverick Capital	Maverick Capital, LTD.	32.7934	-96.8046

Panel A.3: Download by IP=162.216.180.jde on 2016-03-21



Panel B: Summary Statistics of Edgar Downloads

Quarter	Total Machine Download	Total Human Download	Download per Machine IP	Download per Human IP	Daily Machine Download per Firm	Daily Human Download per Firm
2004q1	226,262	112,769	1,859	6	8	8
2004q2	230,506	80,747	2,150	5	8	6
2004q3	75,266	48,156	943	4	5	5
2004q4	129,554	55,961	1,214	5	6	5
2005q1	238,260	85,078	1,818	6	8	6
2005q2	100,933	45,207	1,182	4	6	4
2005q3	69,416	34,245	1,071	6	5	4
2005q4	9,705	3,739	1,985	21	2	3
2006q1	4,866	342	1,627	13	1	2
2006q2	126,993	36,845	3,627	5	3	4
2006q3	179,201	59,183	1,631	5	5	5
2006q4	254,332	60,158	2,177	5	6	5
2007q1	242,608	66,695	1,996	5	7	6
2007q2	225,255	70,943	1,666	5	6	6
2007q3	188,785	65,412	1,517	5	5	6
2007q4	190,639	63,292	1,497	5	6	6
2008q1	255,123	73,150	1,739	5	6	6
2008q2	258,492	71,336	1,522	5	6	6
2008q3	254,237	90,324	1,297	6	8	7
2008q4	409,998	109,779	1,936	6	11	8
2009q1	617,374	143,033	2,509	6	13	9
2009q2	678,892	118,573	2,691	6	8	8
2009q3	833,163	125,985	2,949	6	11	9
2009q4	624,884	145,716	2,248	6	12	10
2010q1	681,293	164,171	2,058	6	12	10
2010q2	1,125,839	155,094	3,319	6	14	10
2010q3	1,400,302	141,788	3,889	6	17	9
2010q4	1,288,033	159,104	3,303	6	13	10
2011q1	1,034,201	202,268	2,159	7	14	12
2011q2	1,055,660	184,611	2,232	7	12	12
2011q3	1,234,590	193,404	2,326	7	11	12
2011q4	1,309,486	168,398	2,523	7	12	11
2012q1	1,677,065	118,983	8,499	7	14	9
2012q2	2,092,169	199,274	2,643	7	16	11
2012q3	2,344,161	201,671	2,555	7	17	9
2012q4	2,565,458	197,254	2,341	6	17	10
2013q1	2,731,777	189,948	2,614	4	16	9
2013q2	3,224,367	261,840	2,111	5	17	12
2013q3	3,329,116	262,708	2,254	5	18	11

Continued on the next page

Continued from the Previous Page

Quarter	Total Machine Download	Total Human Download	Download per Machine IP	Download per Human IP	Daily Machine Download per Firm	Daily Human Download per Firm
2013q4	3,295,713	256,899	2,174	5	21	12
2014q1	3,948,046	300,496	2,517	6	23	12
2014q2	3,627,187	281,838	2,273	5	22	12
2014q3	3,784,509	300,614	2,749	6	26	12
2014q4	3,938,380	270,916	3,054	5	21	11
2015q1	5,078,048	287,422	3,973	5	28	12
2015q2	4,532,023	286,798	3,370	5	26	12
2015q3	5,225,048	276,739	3,977	5	28	11
2015q4	4,419,877	277,478	3,416	4	23	11
2016q1	5,093,468	313,439	3,482	4	25	12
2016q2	5,925,228	345,797	4,049	4	29	14
2016q3	6,974,283	303,755	4,701	4	35	10
2016q4	8,019,323	355,015	4,743	5	37	12

Table 2: Summary Statistics of Key Variables

Table 2 reports summary statistics of key variables. We summarize the definition of the variables in Appendix I. The sample period is 2004-2016.

	Obs	Mean	Std.Dev.	25th Pct.	Median	75th Pct.
Raw Machine Coverage	248,351	2.79	1.28	1.80	2.71	4.01
Raw Human Coverage	248,351	5.24	1.42	4.62	5.38	6.09
Log(Size)	248,351	7.23	1.51	6.07	7.02	8.15
S&P 500	248,351	0.19	0.39	0.00	0.00	0.00
NASDAQ	248,351	0.49	0.50	0.00	0.00	1.00
Log(1+#Filings)	248,351	1.84	0.75	1.10	1.79	2.40
Log(1+#Characters in Filing)	248,351	9.96	3.32	8.89	10.93	12.05
Log(1+#Analyst)	248,351	1.65	1.01	0.69	1.79	2.40
Log(1+#News)	248,351	1.38	1.10	0.00	1.39	2.20

Table 3: Determinants of Machine Coverage

This table presents the result of month-to-month OLS regressions to explain raw machine coverage defined as $\log(1 + Average\ \#Machine\ IP)$. Following Fama-MacBeth (1973), coefficients are calculated as time series averages of monthly estimates, and t-statistics are based on the time series average and standard deviation. Log(Size) is the natural log of the market capitalization, S&P and NASDAQ are dummy variables which are set to 1 when the firm is a member of S&P500 or NASDAQ. Log(1+#Filings) is the natural log of number of SEC filings of the firm in the month. Log(1+#Characters in Filings) is the natural log of the number of total characters in all the SEC filings of the firm in the month. Log(1+#Analyst) is the natural log of (1+number of earning estimates). Log(1+#News) is the natural log of (1+number of media news). Models 1-4 differ in the covariate set used. Model 4 is our baseline model to obtain residual machine coverage. *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

Panel A: Determinants of Machine Coverage

	(1) Machine Coverage	(2) Machine Coverage	(3) Machine Coverage	(4) Machine Coverage
Log(Size)	0.146***	0.132***	0.102***	0.0887***
	(50.08)	(42.50)	(29.32)	(28.13)
S&P500		0.0674***	0.0650***	0.0556***
		(10.81)	(11.00)	(9.51)
NASDAQ		-0.00674*	0.00332	-0.00418
		(-1.94)	(1.15)	(-1.48)
Log(1+#Filings)			0.260***	0.230***
			(16.40)	(14.21)
Log(1+#Characters in Filings)			-0.0140***	-0.0138***
			(-9.98)	(-9.91)
Log(1+Analyst)				0.00988***
				(3.26)
Log(1+News)				0.0513***
				(12.88)
Constant	1.814***	1.904***	1.744***	1.809***
	(24.51)	(24.43)	(18.68)	(18.72)
Observations	248,351	248,351	248,351	248,351
R-squared	0.106	0.108	0.253	0.265

Panel B: Summary Statistics of Adjusted Machine Coverage

	Obs	Mean	Std.Dev.	25th Pct.	Median	75th Pct.
Adjusted Machine Coverage	248,351	1.64	1.23	0.73	1.36	2.86

Table 4: Machine Coverage and Stock Returns: Portfolio Analysis

Table 4 presents average monthly returns for stocks in different quintiles of machine coverage. Average return numbers are presented in basis points. Media coverage of stock *s* in month *t* is estimated as the residual term of the monthly model: $Unajusted\ Machine\ Coverage_{s,t} = \beta_{1,t}Log(Size_{s,t}) + \beta_{2,t}SP500_{s,t} + \beta_{3,t}NASDAQ + \beta_{4,t}Log(1 + \#Filing_{s,t}) + \beta_{5,t}Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t}Log(1 + \#Analyst_{s,t}) + \beta_{7,t}Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$, where Machine Coverage (Unadjusted) of month *t* is estimated as the logarithm of the average number of daily machine IPs which download filings of firm *s* in month t (if there are no machine IPs in a day, then the machine coverage for the stock for the day is 0). The allocation of IP as machines or humans follows the algorithm described in legend to Table 1 (Ryan, 2017). Following Fang and Peress (2009), we then compute the equal-weighted average return of each of the 25 portfolios using individual stock returns in the next month. The monthly returns of each portfolio are either reported raw, or adjusted by DGTW characteristics, or adjusted by risk factors (CAPM, Fama French 3 Factors, Fama French 5 Factors, and Fama French 5 Factor + Carhart Momentum Factor). *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

	Low Machine	2	3	4	High Machine	**Low-High**
1. Raw Returns	116.5***	87.88**	84.26*	86.83*	67.01	49.53***
	(2.85)	(2.06)	(1.93)	(1.95)	(1.46)	(4.32)
2. DGTW Returns	28.03***	-1.352	-5.430	-3.870	-23.46***	51.49***
	(4.85)	(-0.36)	(-1.55)	(-0.88)	(-4.34)	(5.58)
3. CAPM Alpha	43.59***	11.36	6.071	7.100	-14.89	58.48***
	(3.83)	(0.99)	(0.51)	(0.57)	(-1.18)	(5.53)
4. 3 Factor Alpha	44.28***	12.29**	6.930	7.998	-13.94*	58.22***
	(5.70)	(2.03)	(1.17)	(1.20)	(-1.87)	(5.56)
5. 5 Factor Alpha	45.90***	15.89**	7.700	6.819	-12.20	58.10***
	(5.63)	(2.54)	(1.25)	(0.98)	(-1.56)	(5.26)
6. 6 Factor Alpha	47.01***	16.97***	9.220*	7.843	-11.02	58.03***
	(5.94)	(2.86)	(1.69)	(1.18)	(-1.47)	(5.23)

Table 5: Machine Coverage and Stock Returns: Size Quintile Sorts

Table 5 presents average monthly outperformance of stocks with low machine coverage against the high machine quintiles within each size group. Average return numbers are presented in basis points. Each month, we double sort our sample firms machine coverage into five by five quintile groups within each size quintile, and calculate the average outperformance of the lowest machine coverage quintile against the highest qunintile. Media coverage of stock *s* in month *t* is estimated as the residual term of the monthly model: $Unajusted\ Machine\ Coverage_{s,t} = \beta_{1,t} Log(Size_{s,t}) + \beta_{2,t} SP500_{s,t} + \beta_{3,t} NASDAQ + \beta_{4,t} Log(1 + \#Filing_{s,t}) + \beta_{5,t} Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t} Log(1 + \#Analyst_{s,t}) + \beta_{7,t} Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$, where Machine Coverage (Unadjusted) of month *t* is estimated as the logarithm of the average number of daily machine IPs which download filings of firm *s* in month t (if there are no machine IPs in a day, then the machine coverage for the stock for the day is 0). The allocation of IP as machines or humans follows the algorithm described in legend to Table 1 (Ryan, 2017). Following Fang and Peress (2009), we then compute the equal-weighted average return of each of the 25 portfolios using individual stock returns in the next month. The monthly returns of each portfolio are either reported raw, or adjusted by DGTW characteristics, or adjusted by risk factors (CAPM, Fama French 3 Factors, Fama French 5 Factors, and Fama French 5 Factor + Carhart Momentum Factor). *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

	Small	2	3	4	Big
1. Raw Returns	52.03***	67.11***	73.45***	29.02	26.02
	(3.26)	(3.69)	(3.76)	(1.63)	(1.32)
2. DGTW Returns	59.51***	65.01***	71.43***	30.65*	30.86
	(4.43)	(3.97)	(4.01)	(1.93)	(1.64)
3. CAPM Alpha	61.39***	73.42***	82.00***	38.24**	37.36*
	(3.99)	(4.05)	(4.25)	(2.21)	(1.96)
4. 3 Factor Alpha	61.18***	73.27***	82.03***	37.70**	36.89*
	(3.96)	(4.04)	(4.26)	(2.21)	(1.93)
5. 5 Factor Alpha	70.95***	75.98***	75.92***	35.94**	31.71
	(4.44)	(3.97)	(3.76)	(2.01)	(1.58)
6. 6 Factor Alpha	70.18***	78.13***	76.33***	34.79*	30.70
	(4.39)	(4.16)	(3.77)	(1.96)	(1.53)

Table 6: Machine Coverage and Stock Returns: Fama-Macbeth Regressions

This table presents results from a Fama Macbeth (1973) regression of machine coverage on the cross-section of stock returns. Specifically, the econometric model used is as follows:

$$Ret_{s,t} = \beta_0 + \beta_{machine} Machine\ Coverage_{s,t-1} + \beta_1 r_{s,t-1} + \beta_2 r_{s,t-2,t-12} + \beta_3 Volume_{s,t-1} + \beta_4 Log(Size_{s,t-1}) + \beta_5 NASDAQ_{s,t-1} + \beta_6 SP500_{s,t-1} + \beta_7 Log(1 + Analyst_{s,t-1}) + \beta_8 Log(1 + News_{s,t-1}) + \epsilon_{r,t}$$

Here, Machine Coverage is the adjusted machine coverage estimated in the same method as in Table 4. $Ret_{s,t}$ is the stock return of stock s in time *t*, where returns are measured as DGTW adjusted returns, CAPM alphas, Fama French 3 Factor alphas, Fama French 5 Factor alphas and Fama French 5 Factor + Carhart Momentum Factor alphas. $r_{s,t-1}$ is the past month stock return of stock s, and $r_{s,t-1}$ is the cumulative stock return of stock s between [t-12, t-2]. Log(Size) is the natural log of the market capitalization, S&P and NASDAQ are dummy variables which are set to 1 when the firm is a member of S&P500 or NASDAQ. Volume is the monthly trading volume normalized by the number of shares outstanding. Log(1+#Analyst) is the natural log of (1+number of earning estimates). Log(1+#News) is the natural log of (1+number of media news). Following Fama-MacBeth (1973), coefficients are calculated as time series average of monthly estimates, and t-statistics are based on the time series average and standard deviation. *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

	(1) DGTW	(2) CAPM	(3) 3 Factor	(4) 5 Factor	(5) 6 Factor
Machine Coverage, t-1	-96.54***	-95.29***	-90.97***	-88.43***	-88.00***
	(-11.63)	(-11.65)	(-11.89)	(-11.96)	(-12.03)
Ret, t-1	-120.0*	-113.8*	-132.6**	-175.0***	-178.4***
	(-1.90)	(-1.74)	(-2.17)	(-2.91)	(-2.96)
Ret, t-12 to t-2	-7.050	-17.01	-11.77	-6.428	0.830
	(-0.27)	(-0.53)	(-0.40)	(-0.23)	(0.03)
Volume, t-1	-8.766***	-13.70***	-13.40***	-11.81***	-10.64***
	(-2.99)	(-3.98)	(-4.02)	(-3.64)	(-3.51)
Log(Size), t-1	-8.266***	-2.789	-3.160	-3.423	-2.736
	(-3.13)	(-0.45)	(-0.87)	(-0.94)	(-0.76)
NASDAQ, t-1	-4.662	-2.533	-3.529	14.84	11.77
	(-0.52)	(-0.21)	(-0.35)	(1.63)	(1.26)
S&P500, t-1	1.328	2.315	0.123	-1.073	-5.539
	(0.22)	(0.24)	(0.01)	(-0.12)	(-0.62)
Log(1+Analyst), t-1	11.94***	11.90***	10.72***	11.08***	11.07***
	(3.12)	(2.83)	(2.80)	(2.93)	(3.04)
Log(1+News), t-1	-1.012	-0.962	0.400	1.836	1.690
	(-0.33)	(-0.29)	(0.13)	(0.57)	(0.54)
Constant	202.7***	180.7***	179.6***	163.7***	167.8***
	(8.90)	(3.36)	(5.72)	(5.28)	(5.26)
Observations	255,150	255,150	255,150	255,150	255,150
R-squared	0.029	0.046	0.036	0.034	0.032

Table 7: XBRL Implementation, Machine Coverage, and Stock Returns

This table presents results from a difference-in-difference test using a monthly panel regression. Specifically, the econometric model used is as follows:

$$Ret_{s,t} = \beta_0 + \beta_{XBRL\times machine} Post\ XBRL_{s,t} \times Machine\ Coverage_{s,t-1} + \beta_{machine} Machine\ Coverage_{s,t-1} + \beta_{XBRL} Post\ XBRL_{s,t} + \beta_1 r_{s,t-1} + \beta_2 r_{s,t-2,t-12} + \beta_3 Volume_{s,t-1} + \beta_4 Log(Size_{s,t-1}) + \beta_5 NASDAQ_{s,t-1} + \beta_6 SP500_{s,t-1} + \beta_7 Log(1 + Analyst_{s,t-1}) + \beta_8 Log(1 + News_{s,t-1}) + \epsilon_{r,t}$$

Here, $Post\ XBRL_{s,t}$ is an indicator that equals one if the calendar year is after the firm's first XBRL filing on EDGAR. Machine Coverage is the adjusted machine coverage estimated in the same method as in Table 4. $Ret_{s,t}$ is the stock return of stock s in time *t*, where returns are measured as DGTW adjusted returns, CAPM alphas, Fama-French 3 Factor alphas, Fama-French 5 Factor alphas and Fama-French 5 Factor + Carhart Momentum Factor alphas. $r_{s,t-1}$ is the past month stock return of stock s, and $r_{s,t-1}$ is the cumulative stock return of stock s between [t-12, t-2]. Log(Size) is the natural log of the market capitalization, S&P and NASDAQ are dummy variables which are set to 1 when the firm is a member of S&P500 or NASDAQ. Volume is the monthly trading volume normalized by the number of shares outstanding. Log(1+#Analyst) is the natural log of (1+number of earning estimates). Log(1+#News) is the natural log of (1+number of media news). We control for the firm fixed effect and month fixed effect, and estimate the errors by clustering at both firms and months. *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

	(1) DGTW	(2) CAPM	(3) 3 Factor	(4) 5 Factor	(5) 6 Factor
Post XBRL*Machine Coverage, t-1	-0.00335**	-0.00443**	-0.00558***	-0.00494***	-0.00450***
	(-2.19)	(-2.55)	(-3.15)	(-2.85)	(-2.68)
Machine Coverage, t-1	-0.00382**	-0.00313*	-0.00197	-0.00224	-0.00282*
	(-2.55)	(-1.85)	(-1.15)	(-1.33)	(-1.74)
Post XBRL, t-1	0.00491**	0.00626*	0.00962***	0.00895***	0.00856***
	(2.22)	(1.88)	(3.43)	(3.21)	(3.18)
Ret, t-1	-0.0326***	-0.0384***	-0.0389***	-0.0424***	-0.0434***
	(-4.44)	(-4.96)	(-5.68)	(-6.38)	(-6.31)
Ret, t-12 to t-2	-0.00851***	-0.0119***	-0.0116***	-0.0106***	-0.00867***
	(-2.85)	(-3.44)	(-3.58)	(-3.36)	(-2.90)
Volume, t-1	-0.000893**	-0.00101**	-0.000868**	-0.000703*	-0.000925**
	(-2.45)	(-2.26)	(-2.03)	(-1.74)	(-2.42)
Log(Size), t-1	-0.0206***	-0.0225***	-0.0224***	-0.0214***	-0.0203***
	(-17.62)	(-12.80)	(-13.95)	(-12.98)	(-12.89)
NASDAQ, t-1	0.00347	0.00446	0.00366	0.00306	0.00358
	(1.25)	(1.35)	(1.17)	(1.03)	(1.23)
S&P500, t-1	-0.00343*	-0.00180	-0.00242	-0.00217	-0.00224
	(-1.85)	(-0.80)	(-1.11)	(-1.05)	(-1.14)
Log(1+Analyst), t-1	-0.000431	-0.000597	-0.000464	-0.000493	-0.000399
	(-0.81)	(-1.02)	(-0.82)	(-0.89)	(-0.72)
Log(1+News), t-1	-0.000552*	-0.000706*	-0.000563	-0.000494	-0.000476
	(-1.66)	(-1.92)	(-1.61)	(-1.46)	(-1.39)
Observations	230,185	230,185	230,185	230,185	230,185
R-squared	0.028	0.057	0.035	0.035	0.033

Table 8: Determinants of Human Coverage

This table presents the result of the various month-to-month OLS regressions to explain the raw human coverage defined as $\log(1 + Average\ \#Human\ IP)$. Log(Size) is the natural log of the market capitalization, S&P and NASDAQ are dummy variables which are set to 1 when the firm is a member of S&P500 or NASDAQ. Log(1+#Filings) is the natural log of number of SEC filings of the firm in the month. Log(1+#Characters in Filings) is the natural log of the number of total characters in all the SEC filings of the firm in the month. Log(1+#Analyst) is the natural log of (1+number of earning estimates). Log(1+#News) is the natural log of (1+number of media news). Following Fama-MacBeth (1973), coefficients are calculated as time series average of monthly estimates, and t-statistics are based on the time series average and standard deviation. *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

Panel A: Determinants of Machine Coverage

	(1) Human Coverage	(2) Human Coverage	(3) Human Coverage	(4) Human Coverage
Log(Size)	0.374***	0.323***	0.318***	0.296***
	(133.09)	(90.87)	(91.42)	(75.51)
S&P500		0.261***	0.258***	0.248***
		(28.94)	(28.47)	(27.52)
NASDAQ		-0.0195***	-0.0168***	-0.0254***
		(-3.90)	(-3.36)	(-5.05)
Log(1+#Filings)			0.0171*	-0.00676
			(1.91)	(-0.74)
Log(1+#Characters in Filings)			0.0138***	0.0127***
			(7.78)	(6.84)
Log(1+Analyst)				0.0423***
				(8.81)
Log(1+News)				0.0434***
				(9.80)
Constant	2.587***	2.920***	2.765***	2.859***
	(48.45)	(49.71)	(40.90)	(39.81)
Observations	248,351	248,351	248,351	248,351
R-squared	0.212	0.217	0.222	0.228

Panel B: Summary Statistics of Adjusted Human Coverage

	Obs	Mean	Std.Dev.	25th Pct.	Median	75th Pct.
Adjusted Human Coverage	248,351	2.75	1.34	2.19	2.82	3.56

Table 9: Machine Coverage and Human Coverage

This table compares the impact of human coverage and machine coverage on stock returns in a portfolio analysis. Human coverage is the adjusted human coverage ($Human_{s,t}$) of stock s in month t is estimated as the residual term of the monthly model: $Human\ Coverage_{s,t} = \beta_{1,t}Log(Size_{s,t}) + \beta_{2,t}SP500_{s,t} + \beta_{3,t}NASDAQ + \beta_{4,t}Log(1 + \#Filing_{s,t}) + \beta_{5,t}Log(1 + \#Chars\ in\ Filing_{s,t}) + \beta_{6,t}Log(1 + \#Analyst_{s,t}) + \beta_{7,t}Log(1 + \#News_{s,t}) + \beta_{0,t} + \epsilon_{s,t}, \forall t \in T$. Here, human coverage (Unadjusted) of month t is estimated as the logarithm of the average number of daily human IPs which download filings of firm *s* over calendar days in month *t* (if there is no downloading human IPs in a day, then the human coverage of the day is 0). Machine coverage is defined similarly.

Panel A presents the portfolio analysis. In each month, we triple sort our sample firms by human coverage, size and machine coverage into 5 by 5 by 5 groups. We then compute the equal-weighted average return of each of the 125 portfolios using individual stock returns in the next month. Thereafter, we drop the largest size group, and with the remaining size groups, we then take the equal weighted average of portfolio returns over each of Human Coverage and Machine quintiles. 6-factor alphas obtained from applying the Fama-French 5 Factor and the Carhart Momentum factor are presented.

Panel B reports the multivariant Fama Macbeth regressions. Specifically, the econometric model is as follows:

$$\begin{aligned} Ret_{s,t} = \beta_0 &+ \beta_{machine}Machine\ Coverage_{s,t-1} + \beta_{human}Human\ Coverage_{s,t-1} \\ &+ \beta_{machine\ x\ human}Machine\ Coverage_{s,t-1} \times Human\ Coverage_{s,t-1} + \beta_1 r_{s,t-1} + \beta_2 r_{s,t-2,t-12} + \beta_3 Volume_{s,t-1} \\ &+ \beta_4 Log(Size_{s,t-1}) + \beta_5 NASDAQ_{s,t-1} + \beta_6 SP500_{s,t-1} + \beta_7 Log(1 + Analyst_{s,t-1}) + \beta_8 Log(1 + News_{s,t-1}) + \epsilon_{r,t} \end{aligned}$$

Here, Machine Coverage is the adjusted machine coverage, and Human coverage is the adjusted human coverage. $Ret_{s,t}$ represents the stock returns of stock *s* in time *t*, where returns are measured as DGTW adjust returns, CAPM alphas, Fama-French 3 Factor alphas, Fama-French 5 Factor alphas and Fama-French 5 Factor + Carhart Momentum Factor alphas. $r_{s,t-1}$ is the past month stock return of stock *s*, and $r_{s,t-1}$ is the cumulative stock return of stock *s* between [t-12, t-2]. Log(Size) is the natural log of the market capitalization, S&P and NASDAQ are dummy variables which are set to 1 when the firm is a member of S&P500 or NASDAQ. Volume is the monthly trading volume normalized by the number of shares outstanding. Log(1+#Analyst) is the natural log of (1+number of earning estimates). Log(1+#News) is the natural log of (1+number of media news). Following Fama-MacBeth (1973), coefficients are calculated as the time series average of monthly estimates, and t-statistics are based on the time series average and standard deviation. *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

Panel A: Portfolio Analysis						
	Low Machine	2	3	4	High Machine	Low-High Machine
Low Human	67.73***	4.302	-9.368	14.52	7.802	56.72***
	(6.09)	(0.44)	(-0.88)	(1.25)	(0.48)	(3.29)
Mid Human	7.588	19.96*	9.324	3.906	-7.747	15.34
	(0.58)	(1.76)	(0.92)	(0.39)	(-0.72)	(0.93)
High Human	-30.75	24.12	0.917	-15.78	-14.89	-15.86
	(-1.19)	(1.41)	(0.06)	(-1.25)	(-1.19)	(-0.61)
Low-High Human	98.48***	-19.82	-10.28	30.30*	18.67	
	(3.54)	(-0.99)	(-0.53)	(1.83)	(0.90)	

Panel B: Machine Coverage vs Human Coverage: Fama Macbeth Regression					
	(1)	(2)	(3)	(4)	(5)
	DGTW	CAPM	3 Factor	5 Factor	6 Factor
Machine x Human Coverage, t-1	24.95***	22.98***	22.58***	21.52***	20.44***
	(6.65)	(5.39)	(5.44)	(5.22)	(5.09)
Machine Coverage, t-1	-64.77***	-59.31***	-58.17***	-57.85***	-58.45***
	(-5.06)	(-3.95)	(-3.98)	(-4.02)	(-4.43)
Human Coverage, t-1	-42.53***	-43.41***	-39.78***	-37.07***	-36.26***
	(-5.74)	(-5.10)	(-4.98)	(-4.63)	(-5.09)
Ret, t-1	-136.0**	-130.6**	-147.6**	-189.1***	-190.8***
	(-2.16)	(-2.01)	(-2.43)	(-3.16)	(-3.20)
Ret, t-12 to t-2	-12.40	-21.56	-15.94	-10.37	-2.728
	(-0.47)	(-0.68)	(-0.55)	(-0.38)	(-0.11)
Volume, t-1	-9.962***	-14.83***	-14.53***	-12.92***	-11.88***
	(-3.49)	(-4.32)	(-4.38)	(-4.02)	(-3.91)
Log(Size), t-1	-6.318**	-0.989	-1.516	-1.833	-1.149
	(-2.38)	(-0.16)	(-0.42)	(-0.50)	(-0.32)
NASDAQ, t-1	-5.308	-3.070	-4.131	14.30	11.19
	(-0.58)	(-0.25)	(-0.41)	(1.56)	(1.19)
S&P500, t-1	4.010	4.729	2.547	1.129	-3.913
	(0.67)	(0.50)	(0.28)	(0.13)	(-0.44)
Log(1+Analyst), t-1	14.75***	14.60***	13.27***	13.50***	13.38***
	(3.83)	(3.43)	(3.38)	(3.47)	(3.60)
Log(1+News), t-1	-4.096	-3.897	-2.288	-0.779	-0.679
	(-1.29)	(-1.18)	(-0.71)	(-0.24)	(-0.22)
Constant	145.0***	125.9**	126.3***	112.5***	123.4***
	(5.85)	(2.39)	(3.88)	(3.52)	(3.84)
Observations	255,150	255,150	255,150	255,150	255,150
R-squared	0.032	0.050	0.040	0.037	0.036

Table 10: Machine Coverage, Institutional Investors, and Stock Analysts

This table compares the impact of machine downloads in the presence of institutional investors and stock analysts. The former represents the information processing from the buy side, while the later represents the information processing from the sell side. In each month, we triple sort our sample firms by institutional ownership (in Panel A, and analyst coverage in Panel B) , firm size and machine coverage into three by five by five groups. We then compute the equally-weighted average return of each of the portfolios using individual stock returns in the next month. We next drop the largest size group, and with the remaining size groups we compute the equally weighted average of portfolio returns over readability and machine coverage quintiles. The monthly returns of each portfolio are adjusted by a 6-Factor model (Fama French 5 Factors + Carhart Momentum Factor). *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

Panel A: Machine Coverage and Institutional Investors (Buy Side)						
	Low Machine	2	3	4	High Machine	Low-High
Low Institution Ownership	28.21**	6.867	4.958	-13.04	-13.19	41.41***
	(2.37)	(0.60)	(0.48)	(-1.26)	(-1.35)	(2.81)
Mid Institution Ownership	78.66***	34.68***	6.168	26.25***	14.90	63.76***
	(6.54)	(3.28)	(0.61)	(2.69)	(1.04)	(3.41)
High Institution Ownership	61.39***	1.591	-25.11	-31.50*	-19.63	81.03***
	(2.86)	(0.09)	(-1.43)	(-1.94)	(-1.17)	(3.14)
Low-High Institution Ownership	-33.18	5.276	30.06	18.46	6.441	
	(-1.50)	(0.23)	(1.52)	(1.03)	(0.38)	

Panel B: Machine Coverage and Analysts (Sell Side)						
	Low Machine	2	3	4	High Machine	Low-High
Low Analysts	26.14**	2.049	-11.73	-20.58	-8.446	34.59**
	(2.03)	(0.15)	(-0.96)	(-1.60)	(-0.73)	(2.00)
Mid Analysts	78.44***	37.79***	-4.041	4.808	-1.049	79.49***
	(5.41)	(3.10)	(-0.36)	(0.49)	(-0.09)	(4.56)
High Analysts	192.0***	18.51	14.32	-13.94	-1.066	193.0***
	(5.51)	(0.79)	(0.59)	(-0.54)	(-0.04)	(4.31)
Low-High Analysts	-165.8***	-16.46	-26.05	-6.647	-7.380	
	(-4.47)	(-0.64)	(-1.01)	(-0.22)	(-0.24)	

Table 11: Information Readability and Machine Coverage

This table presents the average monthly returns for portfolios constructed by sorting on Machine Coverage and Readability, where Readability is the SEC disclosure text readability estimated as the total characters (in the filings) weighted (Flesch_Reading_Ease + Flesch_Kincaid_Grade_Level + RIX + Coleman_Liau_Index + Gunning_Fog_Index + ARI + SMOG_Index + LIX)/8, where the terms are various readability indices, and each of the indices is standardized by subtracting its monthly mean and divided its monthly standard deviation. In each month, we triple sort our sample firms by readability, firm size and machine coverage into two by five by five groups. We then compute the equally-weighted average return of each of the portfolios using individual stock returns in the next month. We next drop the largest size group, and with the remaining size groups we compute the equally weighted average of portfolio returns over readability and machine coverage quintiles. The monthly returns of each portfolio are adjusted by a 6 Factor model (Fama French 5 Factors + Carhart Momentum Factor). *, **, and *** indicate statistical significance at the 10%, 5%, and 1%, respectively. The sample period is 2004-2016.

	Low Machine	2	3	4	High Machine	Low-High
Low Readability	33.55***	18.45	11.68	-10.61	-3.913	37.46**
	(2.81)	(1.61)	(0.96)	(-0.82)	(-0.30)	(2.31)
Mid Readability	50.40***	12.33	-4.120	8.274	9.830	40.57**
	(3.40)	(0.90)	(-0.37)	(0.79)	(0.91)	(2.42)
High Readability	79.62***	18.80	3.991	10.51	-12.75	92.37***
	(5.75)	(1.42)	(0.36)	(0.91)	(-0.94)	(5.00)
Low-High Readability	-46.07***	-0.350	7.690	-21.12	8.840	
	(-3.06)	(-0.02)	(0.45)	(-1.28)	(0.54)	