Does Big Data Improve Financial Forecasting? The Horizon Effect

Olivier Dessaint, Thierry Foucault, and Laurent Frésard*

September 30, 2020

ABSTRACT

We study how data abundance affects the informativeness of financial analysts' forecasts at various horizons. Analysts produce forecasts of short-term and long-term earnings and choose how much information to collect about each horizon to minimize their expected forecasting error, net of information acquisition costs. When the cost of obtaining short-term information drops (i.e., more data becomes available), analysts change their information collection strategy in a way that renders their short-term forecasts more informative but that possibly reduces the informativeness of their long-term forecasts. Using a large sample of analysts' forecasts at various horizons and novel measures of their exposure to abundant data (e.g., social media data), we provide empirical support for this prediction, which implies that data abundance can impair the quality of long-term forecasts.

Key words: Big data, Financial analysts' forecasts, Forecasting horizon, Forecasts' informativeness, Social media

JEL classification: D84, G14, G17, M41

*INSEAD, HEC Paris, and the Universita della Svizzera Italiana (Lugano), Swiss Finance Institute, respectively. Dessaint can be reached at olivier.dessaint@insead.edu, Foucault can be reached at foucault@hec.fr, Frésard can be reached at laurent.fresard@usi.ch. We thank Randall Morck and participants at INSEAD, the Universita della Svizzera Italiana, and the University of Geneva for useful comments. All errors are the authors' alone. All rights reserved by Olivier Dessaint, Thierry Foucault, and Laurent Frésard.

I Introduction

Progress in computing power and storage infrastructures has triggered an outstanding growth in the volume and variety of data available to the financial industry (e.g., news-feed, social media data, internet traffic data, credit card payments, or satellite images).[1] This evolution transforms how information is produced and used by market participants to predict future outcomes (e.g., cash-flows), make decisions (e.g., choose portfolios) and price assets. Research on its implications for financial markets is still very limited. In particular, the effects of data abundance on the accuracy of investors' forecasts at various horizons are unknown. Yet, understanding these effects is important because many financial decisions require making forecasts at various horizons. For instance, pricing securities or capital budgeting require forming expectations of cash-flows at various points in time in the future.

In this paper, we give a first stab at this issue. We posit that data abundance has reduced the cost of producing information about short-term cash-flows relatively more than about long-term cash-flows. We show theoretically that this shift can induce forecasters to focus relatively more on the production of short-term information, at the expense of the accuracy of their forecasts about long-term cash-flows. Our main contribution is to test this novel prediction and confirm it. Specifically, we find empirically that the emergence of alternative data is associated with a drop in the informativeness of sell-side equity analysts' forecasts about long-term (more than two years) earnings, even though the informativeness of their short-term (less than one year) forecasts improves. This finding is important because financial analysts are central information intermediaries (e.g., Kothari, So, and Verdi (2016)). If data abundance impairs their long-term forecasts, it might negatively affect asset price informativeness and the efficiency of investment decisions.

Progress in information technology reduces the cost of accessing and processing data (e.g., Goldfarb and Tucker (2019) and Veldkamp and Cheung (2019)). However, the cost reduction associated with alternative data is likely to be much stronger for producing short-term information than for producing long-term information. Consider alternative data such as satellite data, credit card data or internet traffic data about a given firm (e.g.,

[1] According to the website AlternativeData.org, there are more than 1500 providers of alternative data in 2020.

satellite images of its parking lots or the number of visits of its website for a retailer). This data clearly contains information about the firm's next quarter earnings but less clearly so about its earnings three years from now.[2] Long-term earnings are likely to be determined by firms' strategic and innovation choices. Predicting the long-term implications of these choices require human judgment and methods of information acquisition that cannot be easily automated (e.g., direct meetings with industry experts, scientists, or managers). Moreover, existing data sources are less likely to be useful to predict these long-term implications (e.g., existing data are unlikely to be useful to understand the potential of radical innovations).

Thus, we posit that data abundance has reduced the cost of producing short-term forecasts of a given accuracy relatively more than the cost of producing long-term forecasts of the same accuracy. To understand the implications of this hypothesis (and ultimately test them), we first consider a forecasting problem in which a financial analyst must forecast both the short-term and long-term earnings of a firm. The long-term earnings is proportional to the short-term earnings plus an orthogonal component (an "*innovation*"), which represents the component of the long-term earnings that cannot be predicted with information about the short-term earnings (e.g., revenues from ongoing investments in innovation).

To form her forecasts, the analyst can collect and process two types of information: (i) information about the short-term earnings ("short-term information") or (ii) information about the *innovation* in the long-term earnings ("long-term information"). With more effort to collect and process information at a given horizon, the analyst obtains a signal of greater precision about the earnings realized at this horizon. We assume that the marginal cost of obtaining a signal (e.g., about the short-term earnings) increases with the precision of this signal (as usual in the literature; e.g., Verrecchia (1982)) *and* the precision of the other signal. This assumption captures the idea that if the analyst puts more effort in sharpening the precision of a signal at a given horizon (e.g., by collecting and processing more short-term information), the cost of increasing further the precision of the other signal increases as well.[3]

[2]For evidence that alternative data contains information about short-term firms' earnings, see Froot, Kang, Ozik, and Sadka (2017), Zhu (2019), Katona, Painter, Patatoukas, and Zeng (2019) and Grennan and Michaely (2019). We are not aware of such evidence for long-term earnings.

[3]For instance, collecting and processing short-term information exhausts cognitive resources of the

The analyst chooses how much effort to devote to the production of short-term and long-term information to minimize her total expected forecasting error (a weighted average of her expected short-term and long-term forecasting errors), net of her total cost of acquiring information. We show that, as the marginal cost of producing short-term information drops, the analyst invests more in obtaining short-term information and less in obtaining long-term information. As a result, the informativeness (i.e., the ability of the forecast to reduce uncertainty about the future earnings) of the analyst's forecast of short-term earnings improves. In contrast, the informativeness of her forecast of the long-term earnings drops if the loss in the accuracy of the analyst's signal about the innovation in the long-term earnings more than offsets the improvement in the accuracy of her signal of the short-term earnings. This happens when (i) the correlation between the short-term and the long-term earnings is low enough so that short-term information becomes less relevant for long-term forecasting, or when (ii) the marginal cost of producing a long-term signal of a given precision increases sufficiently fast with the precision of the short-term signal.

In sum, the model implies that data abundance should increase the informativeness of analysts' forecasts of short-term earnings but can reduce that of long-term earnings. To test this novel prediction, we use a measure of the informativeness of analysts' forecasts at various horizons, which exploits the fact that analysts make recurring earnings' forecasts for multiple stocks at various horizons. Specifically, we measure the overall informativeness of the forecasts of an analyst (a) on a given forecasting day (t) for a given horizon h (ranging from one day to five years) by the R^2 of a regression of realized earnings at horizon h (across stocks covered by the analyst) on the analyst's forecasts of these earnings. A higher R^2 means that her forecasts for horizon h explain (in a statistical sense) a larger fraction of the variation in realized earnings for this horizon, i.e., they are more informative about earnings realized in $t+h$. It also means that the analyst's average squared forecast error *relative* to the dispersion of realized earnings is smaller.[4]

analyst and makes it more costly for her to collect and process additional information, be it short-term or long-term. See Hirshleifer, Levi, Lourie, and Teoh (2019) for evidence of decision fatigue among analysts.

[4]A large mean squared forecast error for an analyst for earnings at a given horizon might stem from the fact that she invests little in information acquisition at this horizon or that prior uncertainty about earnings at this horizon is high (forecasting is more difficult). We are interested in measuring the former effect, not the latter. This is better achieved by using R^2 as a measure of the quality of the analyst' forecast than the mean squared error, although both measures are closely related. To see this formally,

We implement this approach using the earnings' forecasts (horizon up to five years) from I/B/E/S made by 13,465 analysts on 13,436 stocks between 1983 and 2017. Overall, our sample includes more than 65 million analyst-day-horizon observations. We first analyze the relationship between the informativeness of analysts' forecasts and the horizon of these forecasts – the "term-structure of analysts' forecasts informativeness". Perhaps unsurprisingly, and consistent with existing evidence (e.g., Patton and Timmermann (2012) for macro forecasts or van Binsbergen, Han, and Lopez-Lira (2020)), the term-structure of analysts' forecast informativeness has a steep negative slope on average (across all analysts and days). That is, short-term forecasts are significantly more informative than long-term forecasts. For instance, forecasts with horizons shorter than one year explain 79.0% of the variation in realized earnings, compared to 37.62% for forecasts with horizons between three to four years, and 31.18% for horizons comprised between four and five years.

To examine the connection between data abundance and the term-structure of forecasts' informativeness, we first study its time evolution. The amount of digitized data available to analysts has increased over time. Thus, we should observe a "steepening" of the term-structure of analysts' forecasts informativeness over time according to our main prediction. We confirm this prediction. For instance, from before to after 2000 (the middle year in our sample), the informativeness of one-year ahead earnings forecasts increases by roughly 10 percentage points (from about 60% to 70%). In contrast, the informativeness of five-year ahead forecasts drops by roughly 20 percentage points (from more than 40% to less than 30%). In further tests, we formally estimate the annual "slope" of the term-structure of analysts' forecasts informativeness and confirm that this slope has become significantly more negative over time, both in economic and statistical terms. Interestingly, the decline in the informativeness of long-term forecasts relative to short-term forecasts has accelerated in the past decade, which arguably is the period over which the volume of available data has increased the most in our sample.[5]

let the earnings at horizon h be x_h and the analyst's forecast of these earnings be f_{ah}. If these variables are normally distributed, the expected squared forecast error is $EF \equiv \mathsf{E}((x_h - \mathsf{E}(x_h \mid f_{ah})^2 \mid f_{ah})$ and the theoretical R^2 of a regression of x_h on f_{ah} is $R^2_{ah} = 1 - \mathsf{Var}(x_h \mid f_{ah})/\mathsf{Var}(x_h) = 1 - EF/\mathsf{Var}(x_h)$. Thus, R^2_{ah} is higher when the mean squared error of the analyst *relative* to the prior uncertainty about the earnings is higher.

[5]For instance, the volume of new data produced every day has increased from 2 zetabyte in 2010 to 33 zetabytes in 2018 (Statista estimates). Over the same period, the number of alternative data providers and

This evolution is consistent with our main prediction but, of course, it might be driven by many other factors than the growth in the volume of available data, such as changes in analysts' compensation (inducing them to forecast more on short-term forecasts) or increases in uncertainty about long-run earnings (maybe due to the increasing role of innovation in driving these earnings). To address this issue and better isolate the effect of data abundance on the term-structure of analysts' forecast informativeness, we use the introduction and expansion of StockTwits, a large social networking platform where millions of investors share their opinion about individual stocks (e.g., Cookson and Niessner (2020)).

Since its creation in 2009, the number of stocks covered by StockTwits' users has steadily increased and the intensity with which users share information about a stock (measured, for instance, by the number of posts, charts, analyses, or links to articles about a stock) varies greatly across stocks.[6] Importantly, there is evidence that information in blog posts specialized in financial markets (on social medias such as "Estimize", "MotleyFool", "SeekingAlpha"', or "StockTwits",) contains information relevant for predicting short-term stock returns and firms' earnings (e.g., Chen, De, Hu, and Hwang (2014) or Jame, Johnston, Markov, and Wolfe (2016)). Moreover, data vendors such as Bloomberg or Thomson Reuters have gradually integrated StockTwits feed on their terminals for market professionals, which suggests that StockTwits posts help in predicting market moves. In sum, Stocktwits increases the volume of information available to analysts when they form their forecasts, but in a differential way across stocks.[7] For these reasons, it offers an interesting laboratory to analyze how shocks to the volume of data about a stock affects the term-structure of analysts' forecasts informativeness.

We measure the volume of data available on social media for a given stock by the number of StockTwits users having that stock on their "watchlist" or the number of messages exchanged about that stock in the last thirty days. We define the exposure of a

investment in these data by market participants has increased (see https://alternativedata.org/stats/).

[6]For example, Cookson and Niessner (2020) report that a large amount of StockTwits are about Apple and Facebook.

[7]We show in the internet appendix of the paper that analysts are more likely to make a new forecast about a stock following an increase in information produced on StockTwits about this stock, even after controlling for trading activity and the flow of public news about a stock. This finding suggests that analysts use StockTwits as a source of information and are thus likely to be affected by variations in the volume of data available about a stock on StockTwits.

given analyst to social media data by aggregating these measures across all stocks covered by the analyst.[8] We then examine how the informativeness of a given analyst's forecasts at different horizons varies with her exposure to social media data (using analyst and time fixed effects). We find that an increased exposure to social media data is associated with (i) a significant improvement in the informativeness of short-term forecasts, and (ii) a significant drop in the informativeness of long-term forecasts. Thus, consistent with our main prediction, an increase in analysts' exposure to social media data is associated with a significant steepening of the term-structure of their forecasts informativeness.

To support the economic interpretation of this finding, we test three ancillary predictions of our theory. First, the deterioration in the informativeness of analysts' long-term forecasts should be more pronounced when social media data contain more short-term information (so that the marginal cost of producing short-term information drops more). Consistent with this prediction, the negative association between the informativeness of analysts' long-term forecasts and their exposure to social media data is stronger when StockTwits' messages originate from users that self-identify as having short-term horizons (i.e., day-traders and swing traders). Second, we expect the sensitivity of the marginal cost of producing a long-term signal of given precision to the precision of the short-term signal to increase with the number of stocks followed by an analyst. If this is the case, the model implies that the negative association between the informativeness of analysts' long-term forecasts and their exposure to social media data should be stronger for analysts following more stocks. This is indeed the case in our sample. Third, the model predicts that the deterioration of the informativeness of long-term forecast should be stronger when earnings are less auto-correlated because, in this case, information about short-term earnings is less relevant for long-term forecasting. We also find that this prediction holds in our data.

II Related Literature

Our results add to the growing research about the effect of progress in information technology and data abundance on financial markets. Existing studies posit that this evolution

[8] This is similar to Grennan and Michaely (2019) who use the number of messages providing financial analysis of a particular stock in financial blogs as a measure of the production of information by Fintech about this stock.

reduces the cost of accessing and processing information (or relaxes information capacity constraints) and focus on the implications for price informativeness.

For instance, Farboodi and Veldkamp (2020) analyze how investors choose to optimally allocate their limited information capacity between producing information on fundamentals (asset payoffs) or the noise in order flow. They show that as investors' information capacity increases, investors start collecting information about the noise in order flow and may even reduce their effort to produce fundamental information. However, price informativeness increases.

In contrast, Dugast and Foucault (2018) predicts that progress in information technology can reduce price informativeness. In their model, improving the accuracy of a signal takes time so that signals of low accuracy are available before signals of higher accuracy. They show that a reduction in the cost of processing information leads more investors to buy the former signals, which undermines incentives to produce more accurate signals and ultimately lowers price informativeness. Dugast and Foucault (2020) formalize the effect of an increase in the volume of available data on the choice of the precision of their signals by investors, holding the cost of producing signals constant. They show that data abundance can reduce the average quality of investors' signals, even though it improves the quality of the most accurate signals.

Empirically, Zhu (2019) and Grennan and Michaely (2019) find that the introduction of alternative data (such as consumer transactions, satellite images, or financial blog posts) has a positive effect on price informativeness, while Katona, Painter, Patatoukas, and Zeng (2019) find no effect of the availability of detailed satellite images on efficiency.[9]

Our analysis differs in two important ways. First, we analyze how progress in in-

[9]Related research studies how the digitization of financial data affects financial markets. For instance, Gao and Huang (2020) and Goldstein, Yang, and Zuo (2020) study effects associated with the introduction of EDGAR. Since 1993, all public firms in the U.S. must submit various regulatory filings (e.g., forms 10-Ks) electronically on the EDGAR system. This system greatly facilitates investors' access to information about public firms in the U.S. and should therefore reduce the cost of accessing information for investors. Consistent with this possibility, Gao and Huang (2020) find that this introduction is associated with an increase in the informativeness of individual investors' order flow, the number of analysts covering a firm and the accuracy of analysts' short-term forecasts. Goldstein, Yang, and Zuo (2020) find that, following the introduction of EDGAR, firms' investment increases, consistent with a decrease in informational asymmetries between firms and investors. However, they also find a drop in the sensitivity of corporate investment to stock prices, especially for growth firms. They argue that this drop is due to a decline in the production of private information and therefore the informational content of stock prices for firms' managers (RPE).

formation technologies and data abundance affect incentives to produce information at various horizons (the short-term and the long-term). To our knowledge, this question has not been addressed in the literature. Yet, it is important since financial decisions (e.g., asset valuations, portfolio allocations, or capital budgeting) often require making forecasts about fundamental outcomes (e.g., cash-flows) occurring at different dates in the future. Second, we do not focus on price informativeness but rather on the informativeness of analysts' forecasts, who are important information providers. For this reason, our findings also add to the literature on financial analysts, and in particular to recent research focusing on how data abundance is reshaping the financial analysis industry (e.g., Grennan and Michaely (2020)).

III Hypothesis Development

In this section, we present the theoretical framework that guides our empirical analysis of the effects of data abundance on the term-structure of analysts' forecasts' informativeness.

A The Analyst

Figure I presents the timeline of the model. At date 1 an analyst must announce forecasts of the earnings of a firm at different horizons. Specifically, the analyst must forecast: (i) a short-term earnings (e.g., next quarter), θ_{st}, that will be realized at date 2 and (ii) a long-term earnings (e.g., in three years), θ_{lt} that will be realized at date 3. The short-term earnings are normally distributed with mean zero and variance $\sigma_{st}^2 = 1/\tau_{\theta_{st}}$. Long-term earnings are:

$$\theta_{lt} = \beta\theta_{st} + e_{lt}, \tag{1}$$

where e_{lt} is normally distributed with mean zero and variance $\sigma_e^2 = 1/\tau_e$ and independent from θ_{st}. Thus, long-term earnings have two components: (i) one component that depends on short-term earnings and (ii) one component orthogonal to short-term earnings. This specification captures the fact that, in reality, earnings are (positively) autocorrelated. The strength of this correlation in our model is higher if β and $\tau_{\theta_{st}}$ are higher. The component of long-term earnings that is unrelated to short-term earnings can be viewed as being determined, for instance, by R&D investments whose outcomes cannot be predicted by using data informative about short-term earnings.

[Insert Figure I about here]

Let f_{lt} and f_{st} be, respectively, the short-term and the long-term forecasts of the analyst. The analyst' payoff $W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st})$, is realized at date 3, after the realization of the long-term earnings and is inversely related to her short-term and long-term squared forecasting errors:

$$W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st}) = \omega - \gamma(f_{st} - \theta_{st})^2 - (1-\gamma)(f_{lt} - \theta_{lt})^2, \tag{2}$$

where $\omega > 0$ and $\gamma \in [0.5, 1]$. One can interpret $W(f_{lt}, f_{st})$ as the total analyst's compensation from period 2 to period 3 (ω is the maximal compensation). The analyst's payoff is higher if the weighted sum of her unsigned forecasting errors are smaller. The weight γ represents the importance of the short-term forecasting error relative to the long-term forecasting error in determining the analyst's compensation. If $\gamma = 1/2$, the accuracies of the analysts' forecasts about short-term and long-term earnings matter equally for her payoff. If $\gamma > 1/2$, accuracy of her forecasts of short-term earnings matters relatively more. In reality γ will depend on how the analyst's compensation package is designed (i.e., the extent to which this package incentivizes the analyst to produce accurate long-term forecasts), her career concerns (the analyst's overall reputation and ability to increase her compensation might increase with the quality of her long-term forecasts) and discount rates[10].

For given forecasts $\{f_{lt}, f_{st}\}$, the analyst's *expected* payoff at date 1 is:

$$\begin{aligned} \bar{W}(f_{lt}, f_{st}; \Omega_1) &= \mathsf{E}(W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st}))\,|\Omega_1)) \\ &= \overline{W} - \gamma\,\mathsf{E}((f_{st} - \theta_{st})^2\,|\Omega_1) - (1-\gamma)\,\mathsf{E}((f_{lt} - \theta_{lt})^2\,|\Omega_1)\,, \end{aligned} \tag{3}$$

where Ω_1 is the information used by the analyst to formulate her forecasts at date 1.

This information comes from raw data (e.g., accounting data, meetings with the management, industry reports, scientific articles, social media, etc.) that possibly contains (i) information relevant about the common component of short-term and long-term earnings (θ_{st}), and (ii) information relevant about the unique component of long-term earnings (e_{lt}). Thus, after processing all data available to her, the analyst obtains two signals: (i)

[10]Results are identical if the analyst is paid at date 2 based on $(f_{st} - \theta_{st})^2$) and then at date 3 based on $(f_{lt} - \theta_{lt})^2$). In this case, one can interpret an increase in γ as being due to an increase in the discount rate used by the analyst to discount her future wages.

one signal, s_{st} about the common component of short-term and long-term earnings and (ii) one signal, s_{lt} about the unique component of long-term earnings (i.e., $\Omega_1 = \{s_{st}, s_{lt}\}$). We assume that:

$$\begin{aligned} s_{st} &= \theta_{st} + \eta_{st} + \varepsilon_{st}, \\ s_{lt} &= e_{lt} + \eta_{lt} + \varepsilon_{lt}, \end{aligned}$$

where the ηs and the εs are the noise in the analyst's signals. All these noise components are normally distributed and independent from all other random variables in the model (e.g., the firms' fundamentals, θ_{st} and θ_{lt}).

The analyst can reduce the noise coming from ε_{st} and ε_{lt} in her signals by exerting more effort to search, extract and classify information about the common component of short and long-term earnings (henceforth, "short-term information") and the unique component of long-term earnings (henceforth "long-term information"). Effort must be directed and is therefore specific to each horizon. For instance, learning about the unique component of long-term earnings requires collecting and processing data that cannot be used to forecast the common component. To formalize this idea, we assume that $\eta_j \rightsquigarrow N(0, \kappa_j^2)$ and $\epsilon_j \rightsquigarrow N(0, (Z - z_j)\xi_j^2)$, where z_j are the horizon-specific efforts and $j \in \{l, s\}$ $(0 \leq z_j \leq Z)$. In contrast, the analyst cannot learn about the noise coming from the ηs and we assume that $\eta_{jt} \rightsquigarrow N(0, \kappa_{jt}^2)$ for $j \in \{l, s\}$.

We denote by $\tau_j(z_t) = (\kappa_j^2 + (Z - z_j)\xi_j^2)^{-1}$, the precision of signal $j \in \{s, l\}$ for the analyst. The larger is the analyst' effort to collect information about short-term earnings, z_{st}, the higher is the precision of s_{st}, her signal about these earnings. However, this precision cannot exceed $1/\kappa_{st}^2$. That is, even if the analyst devotes maximum effort to learn about short-term earnings ($z_{st} = Z$), her signal remains imprecise. Hence, one can interpret $1/\kappa_{st}^2$ as measuring the extent to which relevant information about short-term earnings is available (if there is a lot relevant information about short-term earnings, $1/\kappa_{st}^2$ is large and the analyst can, with sufficient effort, produce a signal of high quality about short-term earnings). The interpretation of the specification for the long-term signal, s_{lt}, is identical.[11] Last, we note that ξ_{jt}^2 controls both the total uncertainty in the absence of effort ($\tau_{jt}(0)$) and the marginal benefit of one unit of effort (the higher is ξ_{jt}^2, the higher

[11] See **?** for a similar information structure in a different context.

the increase in precision for a one unit effort).

Thus, at date 1, the analysts chooses her forecasts to solve:

$$\text{Max}_{f_{st},f_{lt}} \bar{W}(f_{lt}, f_{st}; s_{lt}, s_{st}). \tag{4}$$

For given efforts z_{st} and z_{lt}, it is easily shown that the analyst's optimal forecasts for short-term and long-term earnings are her conditional expectations of these earnings at each horizon, respectively:[12]

$$\begin{aligned} f^*_{st} &= \mathsf{E}(\theta_{st} \,|s_{st}) \,, \\ f^*_{lt} &= \mathsf{E}(\theta_{lt} \,|s_{st}, s_{lt}) \,. \end{aligned}$$

To simplify the analysis, it is convenient to assume that the analyst has improper priors about θ_{st} and e_{st}. In this case, we have:

$$\begin{aligned} f^*_{st} &= s_{st}, \\ f^*_{lt} &= s_{st} + s_{lt}. \end{aligned} \tag{5}$$

Efforts to collect and process short-term and long-term information is costly. The cost $C(z_{st}, z_{lt})$ of exerting efforts, z_{st} and z_{lt}, to process short-term and long-term information is:

$$C(z_{st}, z_{lt}) = az^2_{st} + bz^2_{lt} + cz_{st}z_{lt},$$

Thus, if $a > 0$ or $b > 0$, the marginal cost of effort ("information processing') increases with the level of effort. This is a standard assumption in the literature on information acquisition (see, for instance, Verrecchia (1982)). We further assume that $c > 0$. It is in line with the two first standard assumptions ($a > 0$ and $b > 0$): If the marginal cost of acquiring a signal increases in its precision then it should logically increase in the precision achieved for other signals. In reality, this captures the idea that if an analyst chooses to put a lot of effort in collecting, say, short-term information then it becomes more demanding for her to make the extra effort of collecting additional information, be it short-term ($a > 0$) or long-term ($c > 0$). For reasons that will become clear below, we

[12]Indeed, these are the forecasts that minimize the short-term and long-term expected squared forecasting error for the analyst)

assume that $4ab > c^2$.

The analyst chooses the efforts she allocates to both short-term and long-term forecasting at date 0, i.e., before generating signals and her forecast to maximize her ex-ante expected payoff net of information acquisition costs. That is, z_{st} and z_{lt} are chosen to solve:

$$\text{Max}_{z_{st},z_{lt}} J(s_{st}, s_{lt}) = \mathsf{E}(\bar{W}(f^*_{lt}, f^*_{st}; s_{st}, s_{lt})), \tag{6}$$

where the analyst's forecasts at date 1, f^*_{lt} and f^*_{st}, are given by eq.(5) (i.e., are chosen optimally). We analyze the solution to this problem and its implication for the informativeness of analysts' forecasts in the next section.

B Optimal Information Acquisition and Forecasts' Informativeness

Using the fact that $f^*_{st} = \mathsf{E}(\theta_{st} \,|s_{st})$ and $f^*_{lt} = \mathsf{E}(\theta_{lt} \,|s_{st}, s_{lt})$, we can rewrite the analyst's objective function at date 0 as:

$$\begin{aligned} J(s_{st}, s_{lt}) &= \overline{W} - \gamma\, \mathsf{E}((f^*_{st} - \theta_{st})^2) - (1-\gamma)\, \mathsf{E}((f^*_{lt} - \theta_{lt})^2), \\ &= \overline{W} - \gamma\, \mathsf{Var}(\theta_{st} \,|s_{st}) - (1-\gamma)\, \mathsf{Var}(\theta_{lt} \,|s_{lt}, s_{st})\,, \\ &= \overline{W} - (\gamma + (1-\gamma)\beta^2))\, \mathsf{Var}(\theta_{st} \,|s_{st}) - (1-\gamma)\, \mathsf{Var}(e_{lt} \,|s_{lt}, s_{st})\,, \end{aligned} \tag{7}$$

where the last line follows from the independence between the short-term component (θ_{st}) and long-term component (e_{lt}) in long-term earnings. Thus, ultimately, the analyst chooses her optimal efforts to minimize the weighted sum of her average unconditional forecasting errors (i.e., average across all realizations of her signals at date 1).

As all variables are normally distributed (and priors are diffuse), we have:

$$\begin{aligned} \mathsf{Var}(\theta_{st} \,|s_{st}) &= (\kappa^2_{st} + (Z - z_{st})\xi^2_{st}), \\ \mathsf{Var}(e_{lt} \,|s_{lt}, s_{st}) &= (\kappa^2_{lt} + (Z - z_{lt})\xi^2_{lt}). \end{aligned}$$

Writing the first-order conditions of the analyst's problem at date 0 (eq.(7)), we deduce

that the analyst's optimal efforts in information acquisition, z^*_{st} and z^*_{lt}, are :

$$\begin{aligned} z^*_{st} &= \text{Min}\{\frac{2bh(\beta,\gamma)\xi^2_{st} - c(1-\gamma)\xi^2_{lt}}{4ab-c^2}, Z\} \\ z^*_{lt} &= \text{Min}\{\frac{2a(1-\gamma)\xi^2_{lt} - ch(\beta,\gamma)\xi^2_{st}}{4ab-c^2}, Z\}, \end{aligned} \tag{8}$$

where $h(\beta,\gamma) \equiv (\gamma + (1-\gamma)\beta^2)$. The second-order condition is satisfied when $4ab > c^2$ (which we assume is the case). Moreover, for values of Z large enough and if $\frac{c\gamma(\beta)}{2a(1-\gamma)} < \frac{\xi^2_{lt}}{\xi^2_{st}} < \frac{2b\gamma(\beta)}{c(1-\gamma)}$, the solution is interior in the sense that the analyst $0 < z^*_j < Z$, for $j \in \{st, lt\}$. Otherwise, at least one of the solution is a corner solution (no effort, $z_j = 0$ or maximal effort, $z_j = Z$). For brevity, in this version of the paper, we focus on the case in which the solution is interior.

We deduce from eq.(8) that the analyst invests more in producing short-term information $\frac{z^*_{st}}{z^*_{lt}} > 1$ if and only if the following condition is satisfied:

$$\frac{\xi^2_{lt}}{\xi^2_{st}} < \frac{\gamma(\beta)(c+2a)}{(1-\gamma)(c+2b)}. \tag{9}$$

Suppose that $\xi^2_{lt} = \xi^2_{st}$, $a = b$ and $\gamma = 1/2$. In this case, the marginal cost and benefit of acquiring short-term and long-term information (in term of improving the precision of the short and long-term signals) are identical for the analyst. Yet, even in this case, the analyst might be more inclined to produce information about the short-term earnings. This can be the case if her payoff depends more on the accuracy of her short term forecast, that is, γ is sufficiently close to one or because the common component of short-term and long-term earnings is sufficiently important relative to the unique component of long-term earnings (i.e., β is large enough). The reason, in the second case, is that the effort to collect short-term information has a greater return since this information can be used to forecast both short-term earnings and long-term earnings.

C Data Abundance and Forecasts' Informativeness

Intuitively, the forecast of the analyst at a given horizon is more informative if observing this forecast enables investors to reduce more their uncertainty about the earnings realized at this horizon. Thus, we define the informativeness of the analyst's forecast at horizon $j \in \{st, lt\}$, denoted by I_j as the inverse of the variance of the firm's earnings at this

horizon conditional on the analyst's forecast at this horizon.[13]. That is:

$$I_j \equiv \mathsf{Var}(\theta_j \mid f_j^*)^{-1} \quad for \quad j \in \{st, lt\} \tag{10}$$

Observe that $\mathsf{Var}(\theta_j \mid f_j^*) = \mathsf{E}((\theta_j - \mathsf{E}(\theta_j \mid f_j^*)^2)$. Thus, the analyst's informativeness at horizon j is larger when her expected forecasting error at this horizon is smaller.

As $f_{st}^* = s_{st}$, we have:

$$I_{st} = \mathsf{Var}(\theta_j \mid s_{st})^{-1} = (\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2)^{-1} \tag{11}$$

Moreover, as $f_{lt}^* = \mathsf{E}(\theta_{lt} \mid s_{lt}, s_{st}) = \mathsf{E}(\theta_{lt} \mid f_{lt}^*)$, we have:

$$I_{lt} = \mathsf{Var}(\theta_{lt} \mid f_{lt}^*)^{-1} = Var(\theta_{lt} \mid s_{st}, s_{lt})^{-1} = (\beta^2(\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2) + \kappa_{lt}^2 + (Z - z_{lt}^*)\xi_{lt}^2)^{-1}. \tag{12}$$

The informativeness of the short-term forecast only depends on the analyst's optimal effort (z_{st}^*) to collect short-term information and naturally increases with this effort. In contrast, the informativeness of the long-term forecast increases in the analyst's efforts allocated to *both* horizons (z_{st}^* and z_{lt}^*) because information about short-term earnings is also useful to forecast long-term earnings when $\beta > 0$.

As explained in the introduction, our hypothesis is that alternative data (e.g., satellite images, social media, mobile phone activity, or credit card transactions) have predominantly reduced the cost of obtaining short-term information, i.e., information relevant for forecasting short-term earnings. In contrast, these data are less useful for forecasting the unique component of long-term earnings. Indeed, this component is more likely to be determined by factors that cannot be easily predicted from alternative data and whose analysis requires expertise and human judgement. In the context of our model, our hypothesis is therefore that data abundance has reduced the cost of producing short-term signals relative to the cost of producing long-term signals. To analyze the effect of this hypothesis, we analyze how a change in a, the factor that determines the rate at which the marginal cost of producing the short-term signal increases with its precision, affects how the analyst chooses the level of her efforts to produce short-term and long-term information, holding other determinants of the total cost of acquiring information (b and c)

[13]This is the same definition as the definition of price informativeness in rational expectations model. See for instance Grossman and Stiglitz (1980)

constant.

Using eq.(8) and focusing on the case in which the analyst's optimal efforts are interior, we obtain:

$$\begin{aligned}\frac{\partial z_{st}^*}{\partial a} &= -\frac{2b}{(4ab-c^2)}z_{st}^* < 0,\\ \frac{\partial z_{lt}^*}{\partial a} &= \frac{2c}{(4ab-c^2)}z_{st}^* > 0.\end{aligned} \tag{13}$$

Not surprisingly, a drop in the marginal cost of obtaining short-term information ("data abundance") leads the analyst to put more effort to improve the accuracy of the short-term signal. Thus, the informativeness of her short-term forecast unambiguously increases with data abundance because:

$$\frac{\partial I_{st}}{\partial a} = (\frac{\partial z_{st}^*}{\partial a})\frac{\xi_{st}^2}{(\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2)^2} < 0. \tag{14}$$

For instance, in the past, it was difficult, if not prohibitively expensive, for analysts to harness the wisdom of crowds to obtain information about future earnings. Now, they can at a low cost follow discussions on social medias about a firm's prospect and use this information as an input in their forecasts of its future earnings (in addition to other, more traditional, sources of information).[14] Even though the cost of accessing to this type of information has dropped, it requires attention from the analyst, which makes the marginal cost of collecting other types of information higher (e.g., it becomes mentally more demanding for the analyst to focus on the firm's long-term prospects after having spent much time to follow discussions about a stock on social medias). In our model, this effect arises when $c > 0$. In this case, as shown by eq.(13), a drop in the marginal cost of obtaining short-term information leads the analyst to reduce her effort to collect long-term information ($\frac{\partial z_{lt}^*}{\partial a} > 0$).

These effects have an ambiguous impact on the informativeness of the analyst' forecast

[14]For instance, a brochure from Deustche Bank emphasizes the usefulness of Estimize (a social media that crowdsources estimates of future earnings from many individuals) to forecast short-term earnings relative to other sources (See "*The wisdom of crowds: crowdsourcing earnings estimates*", Deustche Bank Market Research, March 4 2014. Specifically, it notes that "*Estimize allows individuals to contribute their estimates anonymously. The underlying concept of the community is to capture the wisdom of the crowds in order to reflect investor sentiment and more timely and accurate earnings forecasts*" and notes that one limitation of Estimize is the short-term nature of the forecasts: "*We should also be aware of the potential issues with the Estimize dataset. The main issue rests on [...] the short-term nature of the forecasts"*, in line with our main hypothesis.

for long-term earnings. On the one hand, the analyst collects more short-term information and she can also use this information to improve her forecast of the common component of short-term and long-term earnings. This effect tends to improve the informativeness of her long-term forecast. On the other hand, she collects less information about the unique component of long-term earnings, which tends to reduce the informativeness of her long-term forecast. It is easily shown that the second effect dominates if $\beta^2 < \frac{c\xi_{lt}}{b\xi_{st}}$. Indeed, using eq.(12), we obtain that:

$$\frac{\partial I_{lt}}{\partial a} = (\beta^2 \frac{\partial z^*_{st}}{\partial a}\xi^2_{st} + \frac{\partial z^*_{lt}}{\partial a}\xi^2_{lt})I^2_{lt} = -(\frac{2(\beta^2\xi^2_{st}b - c\xi^2_{lt})}{(4ab - c^2)}z^*_{st})I^2_{lt}. \tag{15}$$

Thus, when $\beta^2 < \frac{c\xi_{lt}}{b\xi_{st}}$, then $\frac{\partial I_{lt}}{\partial a} > 0$. Therefore, a *decrease* in the marginal cost of producing short-term information, a, reduces the informativeness of the analyst's forecast of long-term earnings.

In sum our model has the following prediction:

Main Implication: Data abundance (a drop in a) causes an increase in the informativeness of analysts' short-term forecasts but it can reduce the informativeness of their long-term forecasts.

Note that data abundance reduces the informativeness of the analyst's long-term forecasts when the autocorrelation of firms' earnings is low enough (β small enough), and also when c is large enough. In reality, we conjecture that c should increase with the number of stocks followed by an analyst. Indeed, c captures the fact that if the analyst devotes more attention to another forecasting task than forecasting the unique component of short-term earnings, then her ability to free attention to this task is impaired (e.g., it becomes cognitively more costly to do so). Intuitively, the number of alternative forecasting tasks for an analyst should increase with the number of stocks followed by the analyst and thus c should increase with this number.

Remark: There is evidence that analysts' forecasts are positively biased. This fact does not affect our measure of analysts' forecast informativeness if the bias is a constant (more generally if it does not depend on the signals collected by the analyst). To see this, suppose that after forming her optimal forecast, f^*_j the analyst biases it by a fixed amount B_j (for reasons outside the model). The analyst's reported forecast at horizon j,

denoted f_j^{r*} is then:

$$f_j^{r*} = f_j^* + B_j. \tag{16}$$

Now, as B_j is a constant, we have: $\mathsf{Var}(\theta_j \mid f_j^{r*}) = \mathsf{Var}(\theta_j \mid f_j^*)$. Thus the informativeness of the analyst forecast is not affected by her bias. If the bias is not constant (e.g., a random noise term with positive mean), the analyst's bias reduces the analyst' forecast informativeness but it does not change our comparative static results regarding the effects of exogenous parameters (e.g., a) as long as (i) the analyst's bias is a deviation from the optimal unbiased forecast given the analyst' information (i.e., $\mathsf{E}(\theta_j \mid \Omega_1)$ and (ii) the bias does not depend on the realization of the analyst's signals.

D Testing whether Data Abundance Distorts Forecasts

Designing a test that perfectly identifies the model's prediction requires (i) an adequate empirical proxy for the informativeness of analysts' short-term and long-term forecasts, and (ii) the ability to isolate variation in the relative marginal cost of extracting short-term information from raw data, while maintaining all the other parameters of the model constant. To approximate this ideal setting, we use a large panel of analysts' earnings forecasts over multiple horizons combined with the actual realizations to construct a new empirical proxy for the horizon-specific informativeness of analysts' forecasts that closely maps that in the model.

We then use two distinct settings to capture relative decrease in the cost of extracting short-term information from data. First, we focus on the time-series dimension and study the aggregate evolution of the informativeness of analysts' forecasts at different horizons, conjecturing that data abundance has increased over time. Second, we focus on the cross-sectional dimension, and exploit variation in analysts' exposure to social media data (across analysts' and over time) to capture changes in their marginal cost of extracting short-term information from raw data.

IV Data and Measurements

A Earnings Forecasts and Realizations

We construct a large sample of forecasts' informativeness at different horizons using analyst-by-analyst earnings per share (EPS) and net income forecasts (expressed in US

dollars) from the I/B/E/S Detail History File (Adjusted and Unadjusted). We exclude quarterly and semi-annual earnings forecasts, and retain annual earnings forecasts associated with a clearly defined fiscal period.[15] We eliminate forecasts with missing announcement date, analyst code, or broker code. When a given analyst makes multiple forecasts for a given firm and horizon on a given day, we keep the last forecast based on I/B/E/S time stamp. We further eliminate forecasts that cannot be matched to CRSP, and forecasts made on firms with missing information on stock price, number of shares, and with share code different from 10, 11 or 12. We rely on net income forecast as our main measure of "earnings" forecast. If an analyst makes both a net income and EPS forecast for the same firm and fiscal period on a given day, we retain the net income forecast. If an analyst makes only an EPS forecast, we convert it into a net income forecast. We make that conversion by multiplying the actual net income (see below) by the ratio of the I/B/E/S adjusted EPS forecast over the I/B/E/S adjusted actual EPS. This approach ensures that the implicit number of shares used in the conversion is adjusted for stock splits if needed, and in a way that is consistent with how I/B/E/S adjusts for splits.

Next, we match earnings forecasts to realized earnings reported in the I/B/E/S Actual File (Adjusted and Unadjusted). By default, we use the actual net income to measure realized earnings. When no actual net income is available, but an actual EPS exists, we convert it into actual net income using the fully diluted number of shares from Compustat if the firm does not have multiple shares, and the number of shares from CRSP otherwise. To be included in our final sample of earnings forecasts, we finally apply the following set of criteria. First, all earnings forecasts must be about a fiscal year ending between 1983 to 2017. Second, actual earnings for the forecasted fiscal period as well as total assets from Compustat at the end of the forecasted fiscal period cannot be missing. Third, the earnings forecast must be issued before the actual earnings announcement date, and the actual earnings announcement date must occur after the end of the forecasted fiscal period. To avoid outliers, we finally disregard earnings forecasts that are in absolute value ten times greater than the total assets of the corresponding firm at the end of the forecasted fiscal period.[16]

[15] We identify forecasts for different fiscal years using I/B/E/S item "fpi" and retain forecasts with fpi=1,2,3,4,5,E,F,G,H or I.

[16] For the same reason, we also impose that actual net income (in absolute value) is not greater than

B Measuring Forecasts Informativeness

Measuring the informativeness of a given forecast is challenging, because of the need to estimate a conditional variance for the forecasted variable when one typically observes a single realization of that variable (and not the entire posterior distribution). To overcome this problem, we exploit the fact that, at a given point in time, analysts typically forecast earnings for several firms, providing us with several data points (one for every covered firm) to estimate a conditional variance. We thus rely on the variation in realized earnings and forecasts across the firms covered by a single analyst to compute our informativeness measure.[17] This approach has two important implications. First, our empirical measure of informativeness is analyst-specific only, whereas it is both firm and analyst specific in the model. Our model considers the informativeness of the forecast of an analyst about the earnings of one firm, but empirically, we consider the overall informativeness of the forecasts made by an analyst about several firms. Second, we assume that the underlying distributions of earnings are identical across firms (for the same analyst on the same day at the same horizon), and treat each earnings realization as a realization of the same underlying variable (θ in the model), enabling us to empirically compute the associated conditional variance.[18]

We construct a daily measure of informativeness by analyst and forecasting horizon.[19] The horizon of our measure can vary between one day and five years, depending on whether the analyst discloses earnings forecasts for the current fiscal period, for the next fiscal period, or for subsequent ones. To construct this measure, we use all earnings forecasts most recently issued by an analyst for a specific (future) fiscal period (hereafter the forecasted fiscal period). Specifically, for each analyst and forecasted fiscal period, we create a firm-day panel with all forecasts issued by the analyst for that fiscal period. The panel starts on the date of the first forecast and ends when the covered firms announce

total assets at the end of the forecasted fiscal period.

[17] Our proposed approach is related to Hilary and Hsu (2013) who propose to measure the informativeness of analysts forecasts using the time-series volatility of their errors (i.e., their consistency).

[18] Since analysts tend to follow firms with similar product market characteristics, heterogeneity across firms within analyst is usually low. To make our distribution homogeneity assumption even more realistic, we further normalize (forecasted and realized) earnings of each covered firm by total assets. That is, our assumption is that the distribution of normalised earnings across firms followed by the same analyst is the same.

[19] We build a (high-frequency) daily measure to adequately exploit the granular variation in data abundance in later tests.

their earnings.[20] Every day, the horizon decreases by one day. Each date of the panel is thus associated with a unique horizon measure, defined as the number of days until earnings are disclosed, divided by 365.[21] Since analysts do not update their forecast daily, the panel has gaps, which we fill using the last available forecast whenever it is possible. If no new forecast is issued by the analyst for more than 365 days, the analyst is flagged as inactive from the 366$^{\text{th}}$ day onwards. At the end of this process, a given analyst-day-horizon assembles a collection of forecasts made about various firms for a given forecasting horizon.

We define the informativeness of the forecasts of an analyst (i) on a given day (t) for a given horizon (h) as the R^2 obtained from the following linear regression:

$$e_j = k_0 + k_1 \hat{e}_j + \nu_j, \tag{17}$$

where j indexes all firms covered by analyst i at time t with available forecast at horizon h, and where $\widehat{e_j}$ and e_j are the (normalized) forecasted and realized earnings for firm j, respectively.[22] By definition, the R^2 of this regression is:

$$R^2_{i,t,h} = 1 - \frac{\mathsf{Var}(\nu_j)}{\mathsf{Var}(e_j)} = 1 - \frac{\mathsf{Var}(e_j|\widehat{e_j})}{\mathsf{Var}(e_j)}. \tag{18}$$

Higher $R^2_{i,t,h}$ indicates greater informativeness. Intuitively, when $R^2_{i,t,h}$ increases, the forecasts of analyst i at time t explain a larger fraction of the variation in realized earnings in her portfolio for horizon h, i.e., they are more informative about earnings realized at $t+h$. Notice that $R^2_{i,t,h}$ is the empirical analogue of I_h in our model (with $h \in (st, lt)$). Our empirical measure of informativeness thus maps with the one we use in the theory section (the inverse of $\mathsf{Var}(\theta|f^*)$), with one notable difference. In $R^2_{i,t,h}$, the conditional variance of the forecasted variable is scaled by its unconditional variance. This normalization is important empirically. It ensures that changes in forcasts' informativeness are not counfonded by possible changes in the difficulty to predict earnings, as is the case when

[20] If earnings announcement dates differ across firms, the panel ends on the date of the last earnings announcement.

[21] If earnings announcement dates differ across firms in the panel, we compute the median date and define the horizon as the number of days until that median date.

[22] We normalize both the realized and forecasted earnings by total assets at the end of the forecasted period. We find the same results when normalizing by total assets from the last available financials at t. One drawback of this alternative normalization, however, is that the resulting measure of informativeness sometimes changes even when analysts do not update their forecasts (because the normalization changes).

the variance of earnings changes. In the data, the variance of earnings may vary over time, and in the cross-section of the covered firms. Our use of $R^2_{i,t,h}$ allows us to account for this heterogeneity that is absent in the model.

We obtain $R^2_{i,t,h}$ by estimating regression (17) for each available analyst-day-horizon collection. To be included in the sample, $R^2_{i,t,h}$ must satisfy a number of conditions. First, the number of observations in the underlying regression must be greater than 3 (otherwise the regression cannot be estimated) and lower than thirty (to avoid forecasts issued by teams and not individual analysts). Second, the horizon must be greater than one day and lower than five years. Third, the analyst must not be flagged as inactive. Finally, to limit the effect of outliers coming from lower power in some estimations with few observations, we eliminate $R^2_{i,t,h}$ observations when k_1 is in the first percentile in each tail, and set $R^2_{i,t,h}$ to zero when the estimated k_1 is negative.

This procedure yields a sample containing 62,756,608 analyst-day-horizon observations of R^2, obtained from 13,465 distinct analysts who issued 7.8 million unique forecasts about 13,436 distinct firms with forecasting horizon h ranging between one day and five years.

C The Term-Structure of Forecasts' Informativeness

We present in Table I the summary statistics for our new measure of forecast informativeness, $R^2_{i,t,h}$. Across all horizons, the average informativeness of analysts' forecasts is 68.01%, indicating that the average analyst in the sample makes earnings forecasts that explain 68% of the variation in realized earnings across the firms she covers. We note a substantial variation in forecasts' informativeness across analysts, with a sample standard-deviation of $R^2_{i,t,h}$ of 33.90%. An analyst covers 8.12 stocks on any given day on average (used to estimate regression (17)), ranging between three and thirty. Notably, and perhaps unsurprisingly, the sample includes significantly more short-term than long-term forecasts, as the average horizon is 1.11 years (with a standard deviation of 0.83 years). Two mechanical factors contribute to this asymmetry. First, analysts revise their short-term forecasts more often then their long-term forecasts. Second, in many instances we do not observe the realizations associated with long-term forecasts because firms stay less than 5 years in the sample, or because they disappear before their realization is

observed.[23]

[Insert Table I and Figure II about here]

Table I further presents summary statistics separately for the five forecasting horizons. Confirming the unequal breakdown of observations across horizons, the sample includes more than 33 million observations for the forecasting period of less than one year, compared to about 1.3 million observations for horizons comparised between three and four years. The number of firms covered by an analyst also varies across forecasting horizons, with 8.14 firms covered at the one-year horizon compared to 6.70 for horizons comprised between three and four years.[24]

Remarkably, Table I also reveals that the informativeness of analysts' forecasts varies significantly by horizon. The average forecasts' informativeness is 79.60% for horizons shorter than one year, 59.21% for horizons between one and two years, 49.37% for horizons between two and three years, 37.62% for horizons between three and four years, and 31.18% for horizons between four and five years. Consistent with simple economic intuition and existing evidence (e.g., Patton and Timmermann (2012) who consider macro forecasts), the overall term-structure of forecasts' informativeness is thus downward-sloping.

To better illustrate the shape of the informativeness term-structure, we regress $R^2_{i,t,h}$ on binary variables capturing each (daily) horizon (from one day to five years). Figure II plots the estimated coefficients (together with their 95% confidence intervals) and confirms that forecasts at shorter horizons are significantly more informative than forecasts at longer horizons.

A visual inspection of Figure II suggests that "slope" of the informativeness term-strucrure appears non-trivial. Analysts' forecasts are about two times more informative about realized earnings at the one-year horizon than at the five-year horizon. A linear approximation obtained by regressing $R^2_{i,t,h}$ on the forecasting horizons (with one-year increments of h) and a constant indicates that the slope is approximatively -12 (with a t-statistic of -24). Hence, for the whole sample, the informativeness of analysts' forecasts

[23]The fraction of long-term forecasts also mechanically decreases for all firms after 2015 because we observe realized earnings until 2018 only.

[24]This difference implies that the $R^2_{i,t,h}$ at longer horizons are obtained from estimations that include less number of observations, and may thus be less precise.

deteriorates by about 12 percentage points for each annual lengthening of their forecasting horizon.

V Test#1: Aggregate Evolution

To test whether data abundance makes forecasts more informative at short-horizons and less informative at long-horizons, we first study the aggregate evolution of the term-structure of forecasts' informativeness. Because data has become more abundant (especially in recent years), the model predicts a "steepening" of the term-structure of forecasts' informativeness over time.

Figure III displays the term-structures corresponding to the periods 1983-2000 and 2001-2017. The results visually suggest that the term-structure has indeed become steeper over the second part of our sample, with long-term forecasts becoming markedly less informative after 2000. To formally test whether this change in the term-structure corresponds to a general trend over the sample period, we regress $R^2_{i,t,h}$ on a year counter variable by forecasting horizon sub-sample. This year counter variable is equal to zero before 1992 and increments by one every subsequent year. We further divide this variable by the number of years between 1993 and 2017 so that the estimated coefficient corresponds to the cumulated change in informativeness over the 1993-2017 period.

[Insert Figure III and Table II about here]

We present the results in Table II. Columns (1) and (3) of Panel A, which consider all analysts and no control variables, confirm that the informativeness of short-term forecasts has significantly increased over time. The estimated coefficients on the trend are positive and significant for the forecasting horizons of one year (coefficient of 11.5) and two years (coefficient of 9.4). In sharp contrast, columns (7) and (9) indicate that the informativeness of long-term forecasts has materially deteriorated, with coefficients on the trend of -11.5 for forecasting horizons of four years, and -20 for five years. Confirming the pattern of Figure III, the informativeness of forecasts with forecasting horizon of three years has remained roughly constant over time (see column (5)).

In Panel A, we further report specifications that include fixed effects for two-digit SIC industries (using the main industry covered by each analyst (and year) to assign them into

industries), as well as fixed effects for the average size and age of the covered firms. These inclusions absorb possible changes in forecasts' informativeness stemming from changes in the type of firms covered by analysts. Our conclusions are similar. In Panel B, we further restrict our analysis to analysts issuing forecasts at both short and long horizons, and find a similar shift in the term structure. We conclude that changes in the composition of analysts' portfolios are unlikely to explain the observed steepening of the term-structure of forecasts' informativness.

[Insert Figure IV and Table III about here]

To provide a different perspective on the evolution of the informativeness of forecasts at short and long horizons, we approximate its annual slope by regressing $R^2_{i,t,h}$ on annual increments of h, separately for every calendar year and plot the resulting annual slope coefficients in Figure IV. The slope of the term-structure of forecasts' informativeness has become significantly steeper (i.e., more negative) over time. While the slope remained above -10 until the mid-nineties, its steepening accelerated after 2000. This pattern is confirmed in Table III in which we regress the annual term-structure slope on a normalized trend with annual increments starting in 1993. Column (1) reveals an average slope of -6.6 during the baseline period 1983-1992 (i.e., the estimated constant), followed by a significant steepening after 1993, as the estimated coefficient on the trend is negative (coefficient of -10.6) and statistically significant (t-statistics of -6.26).

The rest of Table III indicates that this conclusion is highly robust. In particular, it holds in columns (2) and (3) when we estimate the slope of the term-structure of forecasts' informativeness for each year and (two-digit SIC) industry. It also holds in columns (4) and (5) when we estimate the slope for each analyst and year (for analyst-year with enough short-term and long-term forecasts). Remarkably, column (5), which reports a specification that includes analysts' fixed effects, shows that the steepening of the informativeness term-structure over time is also present within analyst. This result suggests that the steepening of the informativeness term-structure is unlikely driven by a change in the composition of analysts over time. Finally, Panel B of Table III indicates that our conclusion remains unaffected if we exclude the 80s and focus on the most recent period.

Overall, the steepening of the term-structure of forecasts' informativeness is consistent with the distorting role of data abundance that our theory predicts. Yet, while it seems undeniable that the cost of extracting short-term information from raw data (a) has decreased over time, trends in other forces present in the model could also be consistent with the diverging trajectories of the informativeness of short-term and long-term forecasts that we observe. For instance, a steady increase in investors' focus on accurate short-term forecast (γ) would yield similar predictions. Similarly, a durable decrease in the relative uncertainty about short-term earnings (ξ^2_{st}) compared to long-term earnings (ξ^2_{lt}), perhaps due to improved disclosure or heightened macroeconomic uncertainty, could also be consistent with an aggregate steepening of the term-structure of forecasts' informativeness.

VI Test#2: Exposure to Social Media Data

To better isolate the effect of data abundance on the term-structure of analysts' forecasts informativeness, we focus on one event that contributed to the increase in data available to market participants over recent years: the emergence of social media. Specifically, we study the effect of the introduction, and expansion of StockTwits, a social networking platform for investors where users can publicly share their opinions about stocks and capital markets. Our conjecture is that this increase in publicly available opinions decreased the marginal cost of extracting information about firm short-term cash flows (the parameter a in the model). We first describe StockTwits data, discuss their relevance to test the model's key prediction, and then study the effect of Stocktwits on analysts' forecasts informativeness for different horizons.

A StockTwits Data

StockTwits (www.stocktwits.com) was founded in 2008 as a social networking platform for investors to share their opinions about stocks. The participants can post messages of up to 140 characters and can use $cashtags with stocks' ticker symbols to link their message to particular firms. StockTwits offers its products and services to investors, analysts and the media for the research of stocks and investments.

StockTwits provided data on the universe of messages posted between January 1, 2009 and Decembre 31, 2017. Similar to Cookson and Niessner (2020), we observe for each

message the user identifier, the date, the message content, and the associated $cashtags with the corresponding tickers (a message can be associated with multiple tickers).[25] We also observe specific information about both users and stocks. About users, we have access to self-declared information at the time of registration, including name, location, professional background, investment strategy, and usual investing horizon. For the stocks that users discuss, we know the stock-exchange, and the "watchlist". The "watchlist" for a stock is the number of users who select to follow that stock. For our analysis, we only keep messages about stocks trading on NASDAQ, NYSE, NYSEArca, NYSEMkt, or trading OTC, that are present in CRSP (based on their date and associated tickers) with share code 10,11, and 12. These filters produce a sample containing more than 40 million messages posted by 280,147 unique users about 5,919 unique firms.

[Insert Figure V about here]

Consistent with our conjecture that data have become more abundant in recent years, the intensity of social media activity on StockTwits (i.e., the number of users and their posting intensity) has dramatically increased since its creation, as illustrated in Figure V. For instance, the upper left panel indicates that the number of daily messages increased from about 1,000 in 2009 to about 20,000 in 2013, and 80,000 in 2017. The upper-right panel reveals that the average firm is present in the watchlist of an increasing number of users, rising to about 2,000 in 2017. The lower panels Figure V displays the evolution of the distributions of the daily numbers of messages and watchlists' presence per firm. The figure reveals a substantial and increasing cross-sectional heterogeneity in the availability of social media data across firms. Our tests exploit this time-series and cross-sectional variation.

B Analysts' Exposure to Social Media Data

To examine the relationship between the abundance of social media data and the informativeness of analysts' forecasts at different horizons, we measure their individual exposure to social media data on a given day. To do so, we create two distinct daily measures of data abundance by firm based on the recent activity of StockTwits' users. First, we

[25]Besides Cookson and Niessner (2020), other recent papers also use data from StockTwits to answer different types of questions (e.g., Giannini, Irvine, and Shu (2019) or Cookson, Engelberg, and Mullins (2020)).

consider the total number of users that have that firm on their watchlist on the day before $(t-1)$. Second, we use the total number of messages "cashtagging" that firm in the prior thirty days (from $t-30$ to $t-1$). Similar to Grennan and Michaely (2019), we posit that more users' coverage and messaging activity induces more data available about firms.

To obtain the daily exposure of a given analyst to social media data, we take the average data abundance (i.e., watchlist and messages) across the firms she covers on that day. Hence, an analyst is more exposed to social media data when more data is available (on average) about the firms she covers. For our tests, we consider all analyst-day-horizon in our sample (i.e., with available R^2) between 2005 and 2017, and set analysts' social media exposure to zero when social media information is missing. The resulting sample contains 30,958,705 observations.

[Insert Table IV about here]

Table IV presents summary statistics. The average forecast informativeness of 68.33% is similar to that observed in the whole sample (1987-2017). The forecasting horizon is slightly longer, with an average of 1.26 years (compared to 1.11 in the whole sample), and analysts cover 10.37 firms on average (compared to 8.12). Importantly, the two measures of analysts' exposure social media data display large variability. Firms covered by the average analyst are present in the watchlist of 321 StockTwits users, with a standard deviation 1,471. Similarly covered firms are discussed in 11 messages on average, with a standard variation of 41. The rest of Table IV reports statistics about firm-level variables that we use as controls in our tests. All variables are taken from the last available financials, and aggregated at the analyst-day level (and detailed in the Appendix).

C Relevance Conditions

To adequately capture variation in the marginal cost of producing short-term information from data, our proposed measures of exposure to social media data should satisfy two relevance conditions: Data created by StockTwits's coverage and activity should (i) contain information that is mainly relevant about short-term fundamentals (i.e., earnings), and (ii) used by analysts as part of their source of information (or at least correlated with information used by analysts). We argue that both conditions are likely to hold.

First, existing research indicates that information in social media specialized in financial markets contains information relevant for predicting short-term stock returns and firms' earnings (e.g., Chen, De, Hu, and Hwang (2014), Jame, Johnston, Markov, and Wolfe (2016), Renault (2017), or Bartov, Faurel, and Mohanram (2020)). Although we cannot precisely measure the horizon of information in StockTwits, we rely on the investment horizons of its users as a proxy. Indeed, users can self-declare one of four investment horizon category: "day trader", "swing trader", "position trader", and "long-term investors".[26] Figure VI displays the repartition of messages by users' declared horizons. Consistent with our conjecture that social media data mainly encompasses short-term information, the vast majority of StockTwits' messages stem from users that are either "day traders" (35.4%) or "swing traders". In contrast, only a small fraction of posts are issued by users declaring a long-term horizons, either "position traders" (6.2%) or "long-term investors" (8.6%). We later use this heterogeneity in our tests.

[Insert Figure VI about here]

Second, several arguments suggest that analysts are indeed exposed and sensitive to the information contained in StockTwits' activity. In particular, StockTwits's data has been gradually integrated into all major financial information aggregation platforms commonly used by analysts and other practitioners to source information about firms and industries (e.g., Bloomberg.com, Reuters.com, CNN Money, or Yahoo! Finance, among others). Such integration makes it likely that analysts are exposed to StockTwits' data.

Further consistent with the idea that analysts respond to social media data, we report in the Appendix (to preserve space) several analyses indicating that analysts are significantly more likely to make (or revise) a forecast on a given firm and day following more intense activity by StockTwits' users in the prior thirty days. Remarkably, this result holds when we control for the firm's prior trading volume as well as when we focus only on situations in which there is no news released about firms over the past thirty days (from Capital IQ's key developments data).

Finally, using biographic information on analysts' last names and the first letter of their first names from I/B/E/S between 2009 and 2017 (obtained from the price target dataset),

[26] According to Investopedia.com, "swing traders" have an investment horizon of one or more days, whereas "position traders" have a typicial horizon of several weeks to months.

we find that 35% (of 7,656 distinct analysts) of analysts' names exactly match that of active StockTwits' users (i.e., users that have posted at last one message). Although the matching between I/B/E/S and StockTwits is arguably imperfect (e.g., due to common names), the results nonetheless suggest that a non-trivial fraction of analysts possess StockTwits' accounts and are actively exposed to the resulting social media data.

D Test Specification and Main Results

To assess the role of analysts' exposure to social media data on the informativeness of their forecasts at different horizons, we consider the following baseline specification:

$$R^2_{i,t,h} = \lambda(\text{Social Media Data})_{i,t-1} + \Gamma\text{Controls}_{i,t-1} + \eta_i + \eta_t + \omega_{i,t,h}, \tag{19}$$

where $R^2_{i,t,h}$ is the informativeness of analyst i's forecasts available at time t for the forecasting horizon h. The variable "Social Media Data" is the exposure of analyst i at time $t-1$ on social media data, measured based on StockTwits' watchlist or the number of past messages (define above). The baseline specification includes analysts fixed effects to absorb any time-invariant differences across analysts (e.g., their genuine forecasting ability) and day fixed effects to absorb any variation in forecasts' informativeness that is common across all analysts. We also include control variables capturing characteristics of the firms in analyst i's portfolio that could correlate with the informativeness of her forecasts. Specifically, we consider lagged firms' cash-flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets (inflation adjusted) and the log of age (since their public listing), all aggregated at the level of the corresponding analyst.[27] We cluster the standard errors of $\omega_{i,t,h}$ by forecasted fiscal period.

The coefficient of interest in specification (19) is λ. It measures how, all else equal, temporal variation of an analyst's exposure to social media data (i.e., our proxy for a decrease in the cost of extracting short-term information from raw data) modifies the informativeness of her earnings forecasts for horizon h. Our main prediction is that higher exposure to social media data leads to more informative short-term forecasts (i.e., $\lambda > 0$ for small h) and less informative long-term forecasts (i.e., $\lambda < 0$ for large h). To assess this prediction, we start by estimating the baseline specification (19) separately

[27]Note that, given the fast expansion of StockTwits, we winsorize all variables at the 1% and 99% by date t.

across four distinct groups of horizons (with and without controls), ranging from one year or less ($h \leq 1$) to more than three years ($h \geq 3$).[28]

[Insert Table V about here]

Table V presents the results, with Panel A considering social media data based on the number of users with covered firms in their watchlist, and Panel B based on the number of prior messages. To facilitate economic interpretation, we standardize the variable "Social Media Data" by its standard deviation measured in 2017, which better reflects current variation in social media data available about firms. Across both panels, the first two columns show that the coefficient on "Social Media Data" is positive and statistically significant. More social media data available for the average analyst leads to more informative forecasts at horizons shorter than one year ($h \leq 1$). Columns (3) and (4) indicate that variation in data abundance does not significantly affect analysts' informativeness at mid-term horizons ($1 < h \leq 2$). In sharp contrast, columns (5) to (8) indicate that increased exposure to social media data leads to a significant deterioration of long-term forecasts' informativeness. The estimated coefficients on "Social Media Data" are negative and statistically significant for horizons comprised between two and three years ($2 < h \leq 3$) and longer than three years ($h \geq 3$).

The economic magnitude of the estimated associations appears non-trivial. Across both panels, a one standard deviation increase in analysts' exposure to social media data leads to a deterioration of their informativeness of long-term forecasts between 3.00% to 4.83%, and an improvement in the informativeness of their short-term forecast between 1.03% and 1.62%. In relative terms, the estimated degradation of long-term forecasts' informativeness is about three times larger than the corresponding improvement in the short-term (e.g., coefficients of -4.83 and -3.00 in columns (8) compared to 1.57 and 1.03 in columns (2)).

[Insert Table VI about here]

To provide a different perspective on economic magnitude, we modify the baseline specification (19) by pooling together analyst-day-horizon observations across all hori-

[28]For this test, we group together horizons between three and five years because we have few observations at long horizons.

zons, and include an interaction term between "Social Media Data" and the (annualized) forecasting horizon of each observation (centered at a one-year horizon for convenience). We present the results in Table VI. Confirming the results in Table V , column (1) and (4) reveal that the coefficients on the interaction term are negative and statistically significant with both proxies for social media data. Greater exposure to social media data makes the term-structure of informativeness by analyst steeper. Column (1) (respectively column (2)) indicates that, for a given increase in social media data (e.g., a one-standard deviation increase), the informativeness of analysts' forecasts decreases more than is usually the case. For each annual lengthening in their forecasting horizons (e.g., from $h = 1$ to $h = 2$), informativeness decreases by an extra 3.20% (2.36%) compared to the baseline decline of 16.66% (16.59%) that the downward sloping term-structure typically predicts absent social media exposure.

Remarkably, the rest of Table VI indicates that the relative deterioration of long-term forecasts' informativeness continues to hold when we focus specifically on the variation of the informativeness of the analysts' forecasts within a given annual forecasting horizons (with the inclusion of analyst×forecasting horizon fixed effects). It also holds when we further include date×horizon fixed effects, which absorbs any common variation in the informativeness of the forecasts issued on a given day and for a given horizon.

E Additional Predictions and Ancillary Results

To further document the economic channel at play in the model we test three ancillary predictions of our theory. First, the deterioration in the informativeness of analysts' long-term forecasts should be more pronounced when social media data contains more short-term information (i.e. when the marginal cost of producing short-term information a decreases). Second, the negative association between the informativeness of long-term forecasts and social media data abundance should increase when analysts follow more stocks (i.e. when c is more negative). Third, the negative effect of data abundance on the informativeness of the analyst' long term forecast should be stronger when earnings are less auto-correlated (i.e. when β is low).

Providing support for all these ancillary patterns is important as they lend further support to the idea that the distortion of the term-structure of forecasts informativeness

is indeed due to analysts' reliance on social media data and the resulting shift in their information collection strategy. Any alternative story must explain not only our main finding - a *simultaneous* increase in the informativeness of short-term forecasts and a decrease in that of long-term forecasts - but also all of the specific predictions.

E.1 Users' Investing Horizon (a)

Our model predicts that the deterioration of long-term forecasts' informativeness arises because social media data mainly contains information that is relevant about short-term earnings. To provide support for this premise, we exploit the heterogeneity in investing horizon across StockTwits' users. We posit that users who define themselves as "day traders" are more likely to produce information about the short-term than those who define themselves as "long-term investors". We thus count the number of messages posted over the last thirty days by each category of trader (i.e., "day trader", "swing trader", "position trader" and "long-term investor") for each stock an analyst covers and next compute the average by analyst at time $t-1$. We then use these four variables (instead of the total aggregated number of messages) interacted with the (annualized) forecasting horizon, similar to the specification reported in Table VI.

[Insert Table VII about here]

Table VII reveals that the deterioration of the informativeness of analysts' long-term forecasts related to social media data is only present when the data is produced by users exhibiting short-term horizons. Indeed, the interaction terms between the number of messages and the forecasting horizon is significantly negative only for message written by "day traders" (coefficients ranging between -0.87 and -1.06) and "swing traders" (coefficients ranging between -0.88 and -0.97). Social media data displays no significant relationship with the informativeness of forecasts at different horizons when the data is produced by "position traders" and "long-term investors". Results in Table VII support our hypothesis that the distorting role of data abundance for long-term forecasting stems from the preponderance of short-term information in social media data, and the resulting reallocation of analysts' data processing efforts towards short-term information.

E.2 Analysts' Processing Capacity (c)

Our model also predicts greater distortion of the term-structure when the multiplicity of analysts' tasks increases (i.e. when c is more negative). We consider the number of firms covered by each analyst (and day) as a proxy for task multiplicity. We posit that analysts covering more firms exploit more of their overall data processing capacity, and are thus more subject to processing fatigue. For these analysts, the substitution between efforts to process short-term and long-term information is more costly (i.e., c is more negative).

[Insert Table VIII about here]

To assess whether the distortion of the term-structure of informativeness related to data abundance is larger for analysts that are more prone to processing fatigue, we interact the number of firms they cover with the interaction term between the two measures of social media data and the forecasting horizon (as in Table VI). The resulting triple interaction terms measures whether the worsening of the informativeness of long-term forecasts (i.e., the interaction between "Social Media Data" and horizon) is more pronounced for analysts covering more firms. Table VIII reveals that this is the case, confirming that the condition $\beta^2 b < c$ is more likely to hold when analysts cover more firms. Across various specifications, all coefficients on the triple interaction are negative, and five out of six are statistically significant.

E.3 Correlated Earnings (β)

Finally, we consider the role of β, the parameter governing the temporal link between long-term and short-term earnings. When β is low, the model predicts that the deterioration of the informativeness of long-term forecast should be stronger because information about short-term earnings is less relevant for long-term forecasting. We use firms' earnings auto-correlation as an empirical proxy for β. We obtain it by regressing firms' quarterly earnings on its lag (without constant) using a rolling window of two years (and requiring at least four observations). We then aggregate this measure at the level of analyst (and day) by taking the average auto-correlation across the firms she covers (on that day).

[Insert Table IX about here]

Using the same triple interaction approach as above, Table IX reveals that the negative impact of social media data on the informativeness of long-term forecasts is less pronounced for analysts covering firms whose earnings are more persistent. The coefficients on the triple interactions are all positive and statistically significant. Consistent with the economic mechanism featured in the model, the steepening of the term-structure of informativeness is weaker when earnings are less auto-correlated.

VII Conclusion

This paper examines how data abundance affects the informativeness of financial analysts' forecasts at various horizons. We posit that data abundance has reduced the cost of producing information about short term cash-flows relatively more than about long-term cash-flows. We show theoretically that this shift can induce forecasters to focus relatively more on the production of short-term information, at the expense of the accuracy of their forecasts about long-term cash-flows. Our main contribution is to test this novel prediction and confirm it. Specifically, we find empirically that the emergence of alternative data is associated with a drop in the informativeness of sell-side equity analysts' forecasts about long-term (more than three years) earnings, even though the informativeness of their short-term (less than one year) forecasts improves. If data abundance impairs their long-term forecasts, it might negatively affect asset price informativeness and the efficiency of investment decisions.

References

Bartov, Eli, Lucile Faurel, and Partha Mohanram, 2020, Can twitter help predict firm-level earnings and stock returns?, Working Paper.

Chen, Hailang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, *Review of Financial Studies* pp. 1367–1403.

Cookson, Anthony, and Marina Niessner, 2020, Why don't we agree? evidence from a social network of investors, *Journal of Finance* 75, 173–228.

Cookson, Tony, Joey Engelberg, and William Mullins, 2020, Does partisanship shape investor beliefs? evidence from the covid-19 pandemic, Review of Asset Pricing Studies (forthcoming).

Dugast, Jerome, and Thierry Foucault, 2018, Data abundance and asset price informativeness, *Journal of Financial Economics* pp. 367–391.

——— , 2020, Equilibrium data mining, data abundance and the, Working Paper.

Farboodi, Maryam, and Laura Veldkamp, 2020, Long rung growth of financial data technology, *Forthcoming in the American Economic Review.*

Froot, Kenneth, Namho Kang, Gideon Ozik, and Ronnie Sadka, 2017, What do measures of real-time corporate sales say about earnings surprises and post-announcement returns?, *Journal of Financial Economics* pp. 143–162.

Gao, Meng, and Jiekun Huang, 2020, Informing the market: The effect of modern information technologies on information production, Review of Financial Studies (forthcoming).

Giannini, Robert, Paul Irvine, and Tao Shu, 2019, The convergence and divergence of investors' opinions around earnings news: Evidence from a social network, *Journal of Financial Markets* pp. 94–120.

Goldfarb, Avi, and Catherine Tucker, 2019, Digital economics, *Journal of Economic Literature* 57, 3–43.

Goldstein, Itay, Shiijie Yang, and Luo Zuo, 2020, The real effects of modern information technologies, *Working paper, NBER.*

Grennan, Jillian, and Roni Michaely, 2019, Fintechs and the market for financial analysis, *Forthcoming Journal of Financial and Quantitative Aanalysis.*

——— , 2020, Artificial intelligence and the future of work: Evidence from analysts, working paper.

Grossman, Sanford, and Joseph Stiglitz, 1980, On the impossibility of informationally ef?cient markets, *American Economic Review* pp. 393–408.

Hilary, Gilles, and Charles Hsu, 2013, Analyst forecast consistency, *Journal of Finance* pp. 271–297.

Hirshleifer, David, Yaron Levi, Ben Lourie, and Siew Hong Teoh, 2019, Decision fatigue and heuristic analyst forecasts, *Journal of Financial Economics* pp. 83–98.

Jame, Russell, Rick Johnston, Stanimir Markov, and Michael Wolfe, 2016, The value of crowdsourced earnings forecasts, *Journal of Accounting Research* 54, 1077–1109.

Katona, Zsolt, Marcus Painter, Panos N. Patatoukas, and Jean Zeng, 2019, On the capital market consequences of alternative data: Evidence from outer space, Working Paper.

Kothari, S.P., Erik So, and Rodrigo Verdi, 2016, Analysts' forecasts and asset pricing: A survey, *Annual Review of Financial Economics* 8, 197–219.

Patton, Andrew, and Allan Timmermann, 2012, Forecast rationality tests based on multi-horizon bounds, *Journal of Business and Economics Statistics* 30, 1–17.

Renault, Thomas, 2017, Intraday online investor sentiment and return patterns in the u.s. stock market., *Journal of Banking and Finance* 84, 25–40.

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2020, Man vs. machine learning: The term structure of earbnings expectations and conditional biases, *Working paper, NBER*.

Veldkamp, Laura, and Laura Cheung, 2019, Data and the aggregate economy, *Working paper*.

Verrecchia, Robert, 1982, Information acquisition in a noisy rational expectations economy, *Econometrica* pp. 1415–1430.

Zhu, Christina, 2019, Big data as a governance mechanism, *Review of Financial Studies* 32, 2021–2061.

Figure I: Timeline of the model

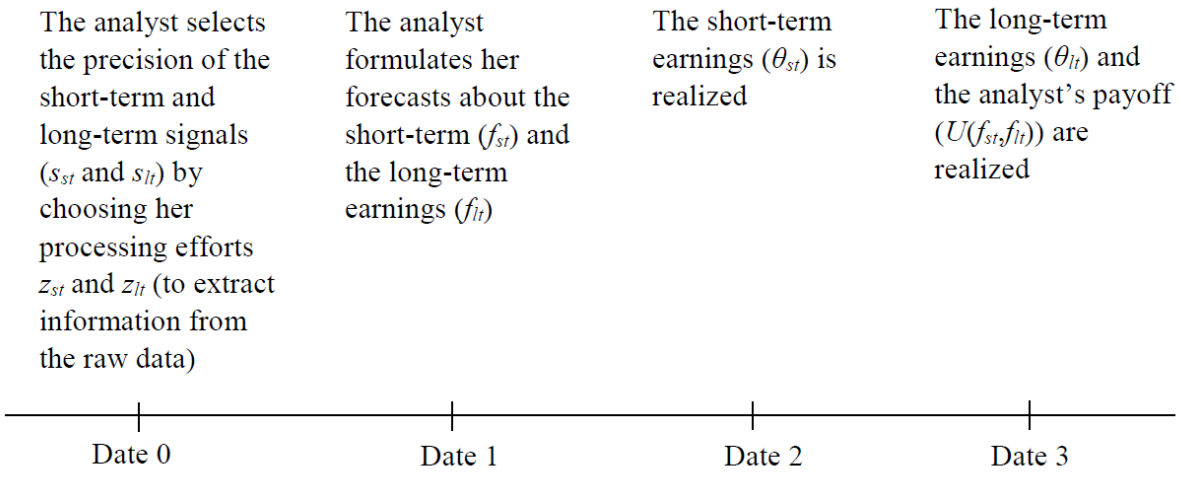
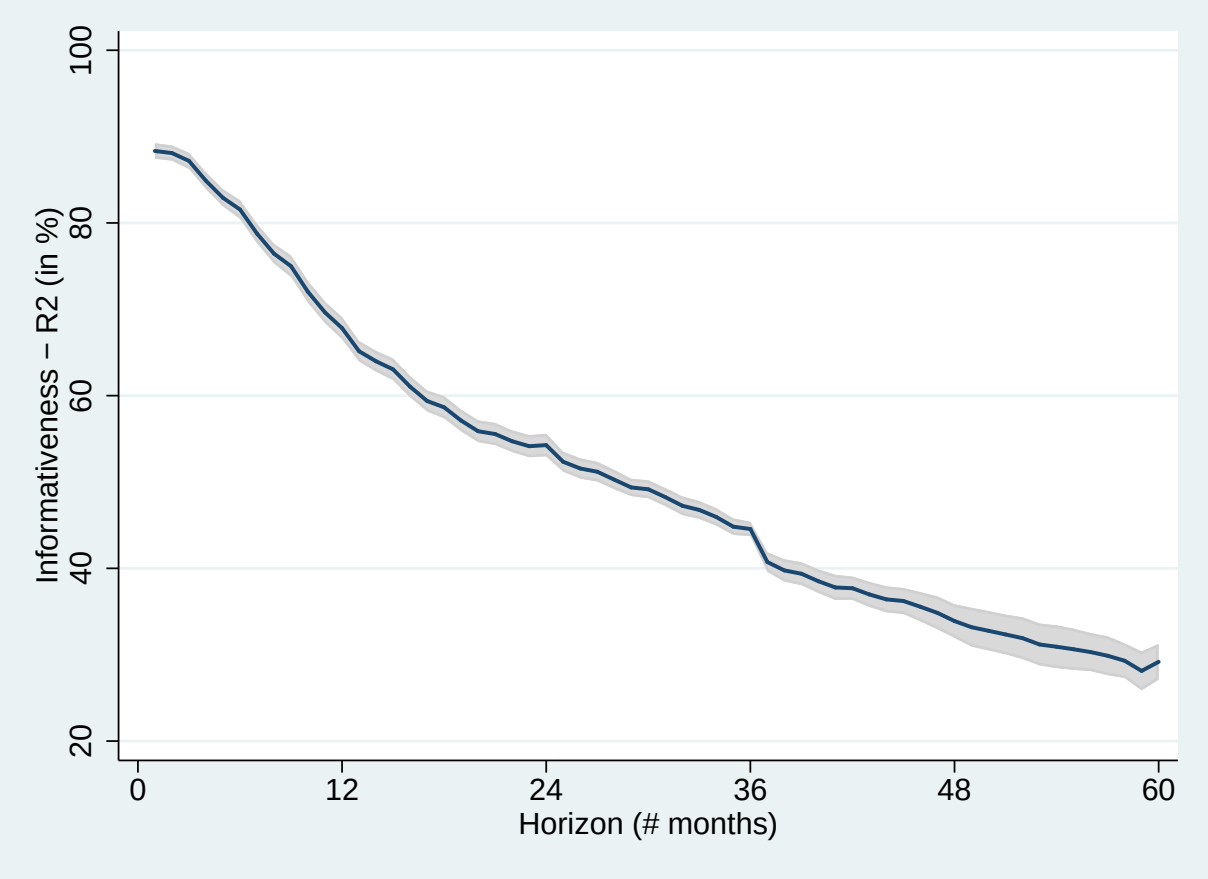
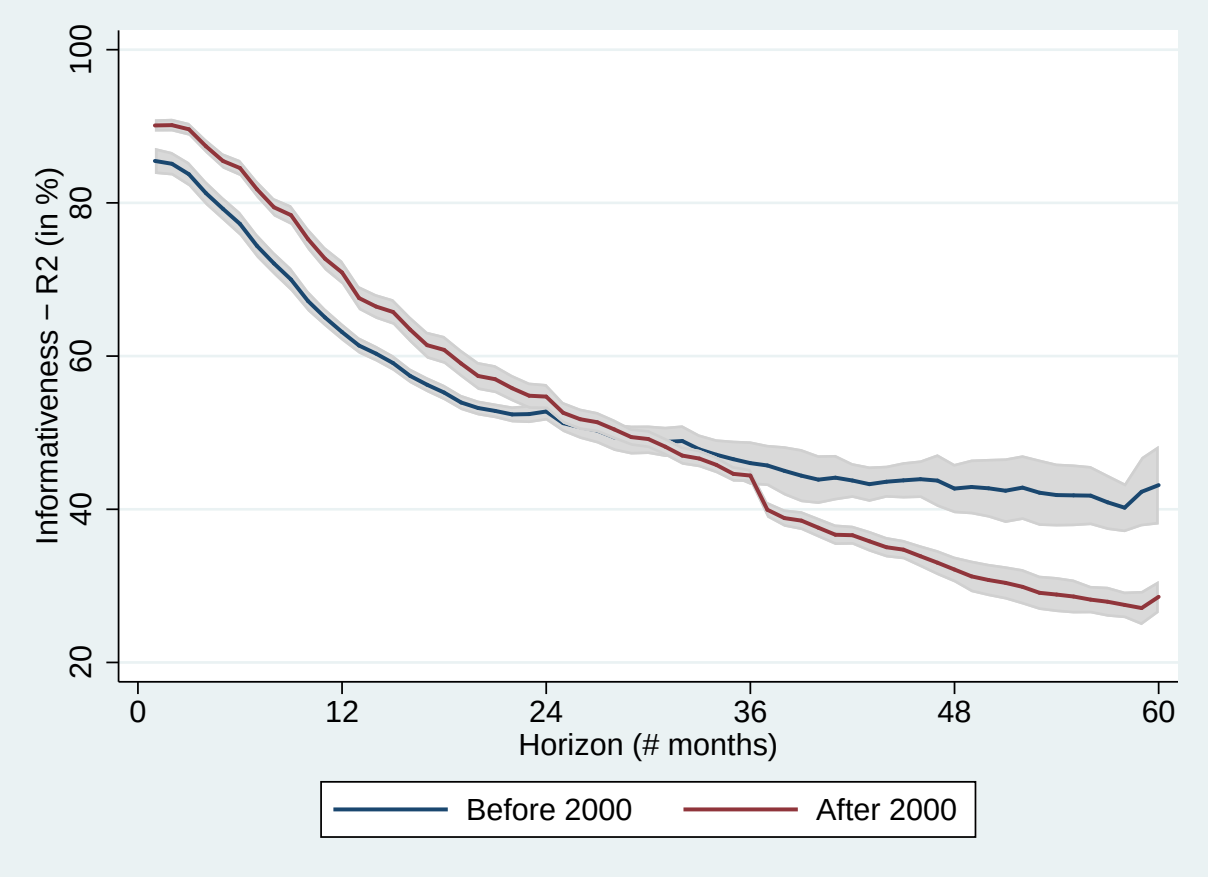


Figure II: The term-structure of analysts forecasts' informativeness



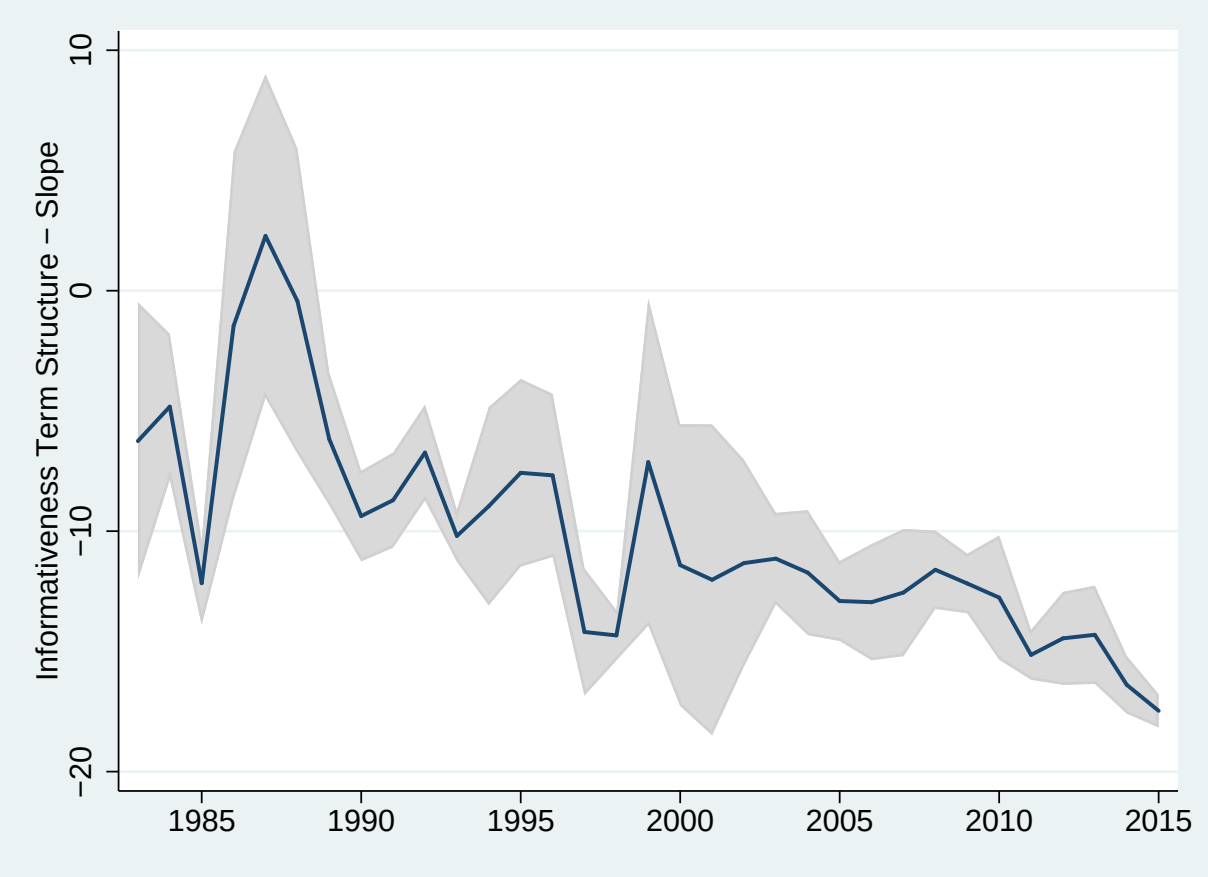
This figure displays the term-structure of analysts forecasts' informativeness. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in months) from zero to five years. The forecasting horizon is measured as the number of days between the forecasting date and the date of actual earnings release divided by 365. The sample period is 1983-2017. The shaded gray area corresponds to a 90% confidence interval.

Figure III: The term-structure of analysts forecasts' informativeness over time



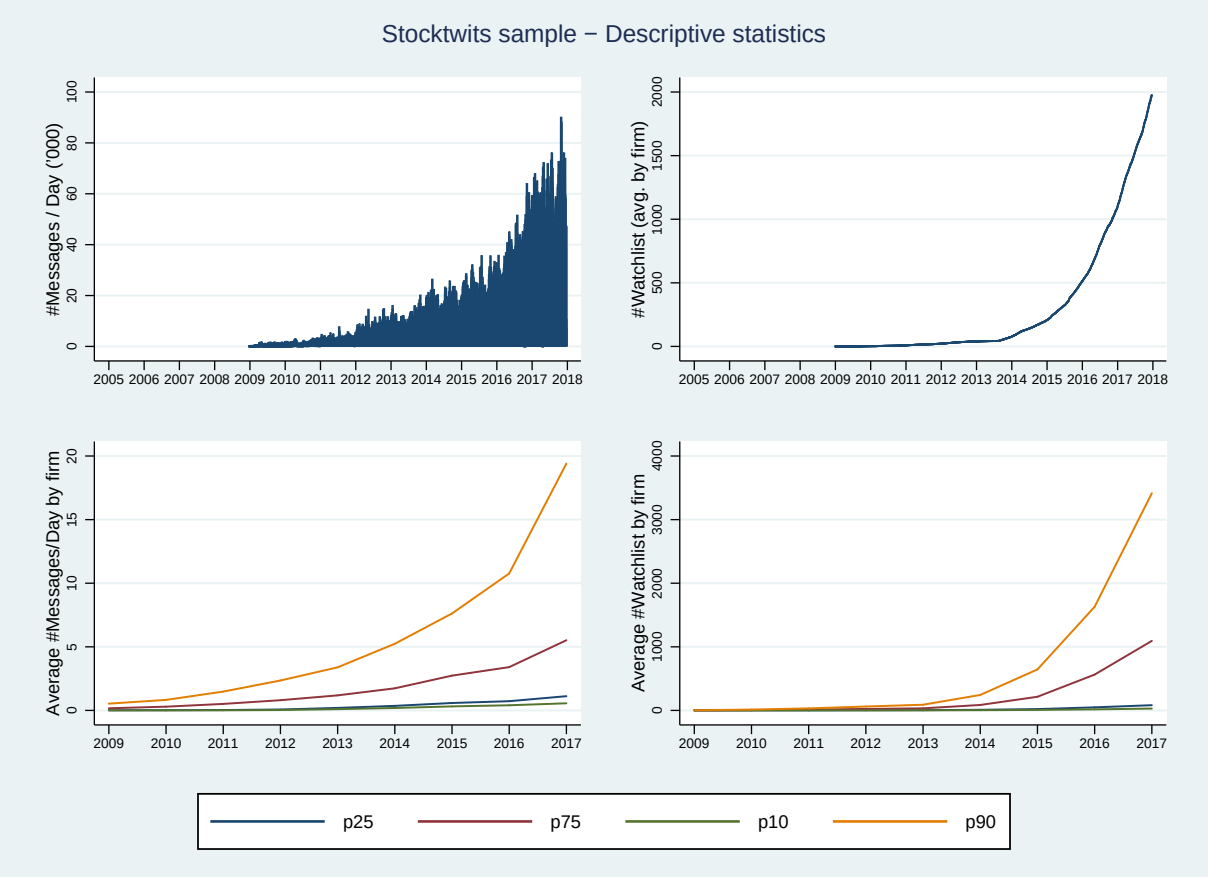
This figure displays the term-structure of analysts forecasts' informativeness before and after 2000. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in months) from zero to five years. The forecasting horizon is measured as the number of days between the forecasting date and the date of actual earnings release divided by 365. The sample period is 1983-2017, split into two sub-period of equal length. The shaded gray area corresponds to a 90% confidence interval.

Figure IV: The slope of term-structure of analysts forecasts' informativeness



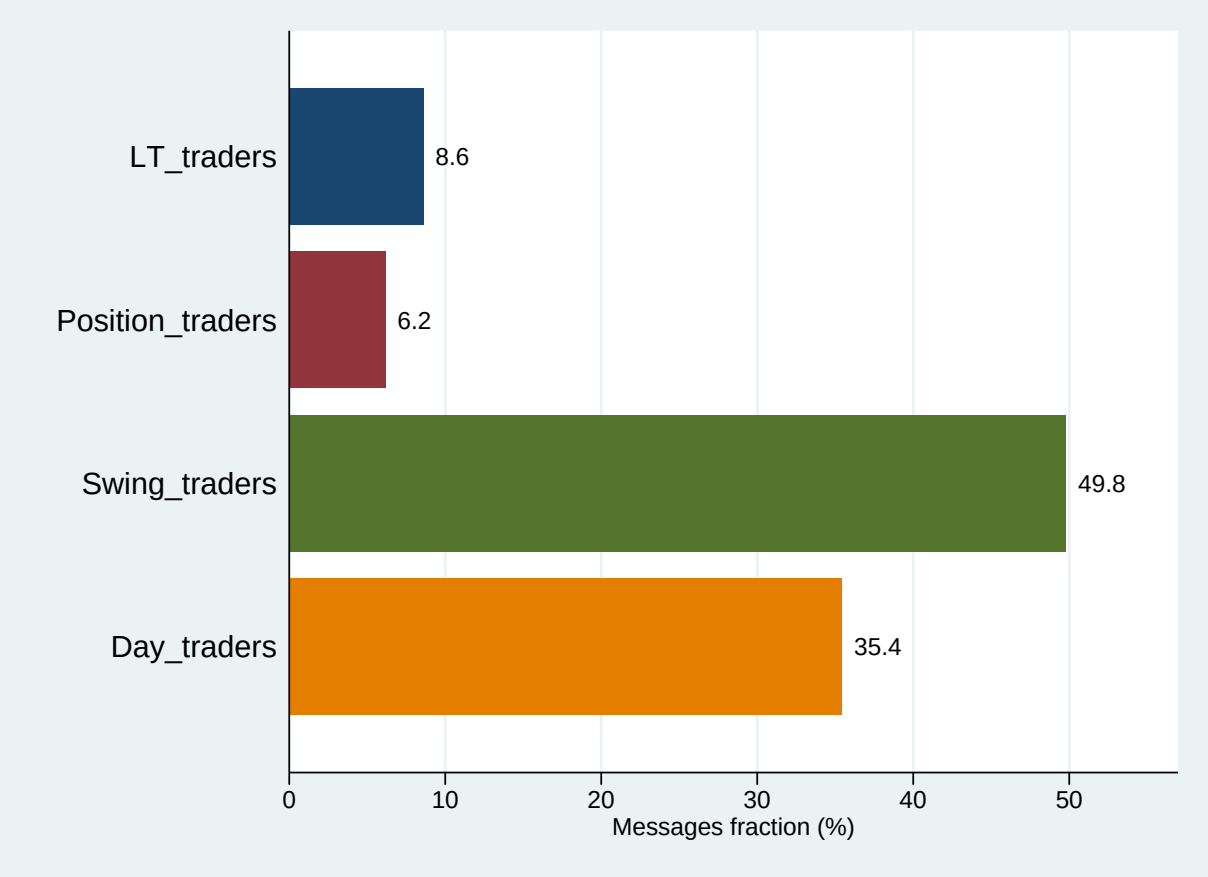
This figure displays the evolution of the slope of the term-structure of analysts forecasts' informativeness. The annual slopes are obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on annual increments of horizon (measured as the number of days between the forecasting date and the date of actual earnings release divided by 365), separately for every calendar year. The figure plots the resulting annual slope coefficients. The shaded gray area corresponds to a 90% confidence interval.

Figure V: StockTwits' Expansion and Social Media Data



This figure displays descriptive statistics on the evolution of StockTwits between 2005 and 2017 (in our sample). The upper-left panel presents the total number of messages per day. The upper-right panel presents the average number of users that have a given firm in their watchlist. The bottom-left panel presents different percentiles of the average number of messages per day and firm. the bottom-right panel presents different percentiles of the average number of users that have a given firm in their watchlist.

Figure VI: StockTwits' users investment horizon



This figure displays the repartition of messages by StockTwits' users declared investment horizons, split into four distinct categories: "day trader", "swing trader", "position trader", and "long-term investors". The sample period is 2009-2017.

Table I: Descriptive statistics

This table presents descriptive statistics for the main analyst-day-horizon variables used in the aggregate tests. R^2 measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. The forecasting horizon is measured as the number of days between the forecasting date and the date of actual earnings release divided by 365. #Stocks is the number of stocks covered by an analyst on a forecasting day. The sample covers the period from 1983 to 2017. We present statistics for the whole sample, as well as sub-samples including observations in different forecasting horizon ranges. Detailed variable definitions are provided in the Appendix.

	N	Mean	St.Dev	Min	P25	P50	P75	Max
				Whole sample				
R^2	65,888,460	68.01	33.90	0.00	45.71	82.70	96.30	100.00
horizon	65,888,460	1.11	0.83	0.00	0.48	0.99	1.56	5.00
#Stocks	65,888,460	8.12	5.18	3.00	4.00	7.00	11.00	30.00
				Sample: horizon <= 1 Yr				
R^2	33,413,667	79.60	27.63	0.00	72.57	92.49	98.42	100.00
horizon	33,413,667	0.49	0.29	0.00	0.24	0.49	0.74	1.00
#Stocks	33,413,667	8.29	5.36	3.00	4.00	7.00	11.00	30.00
				Sample: 1 Yr <= horizon <2 Yrs				
R^2	25,060,925	59.21	34.64	0.00	29.37	69.51	90.42	100.00
horizon	25,060,925	1.45	0.28	1.00	1.21	1.43	1.68	2.00
#Stocks	25,060,925	8.14	5.09	3.00	4.00	7.00	11.00	30.00
				Sample: 2 Yrs <= horizon <3 Yrs				
R^2	5,361,069	49.37	36.23	0.00	10.47	53.15	84.34	100.00
horizon	5,361,069	2.39	0.28	2.00	2.15	2.34	2.61	3.00
#Stocks	5,361,069	7.53	4.71	3.00	4.00	6.00	10.00	30.00
				Sample: 3 Yrs <= horizon <4 Yrs				
R^2	1,349,749	37.62	36.04	0.00	0.00	28.84	71.60	100.00
horizon	1,349,749	3.45	0.29	3.00	3.20	3.43	3.70	4.00
#Stocks	1,349,749	6.70	3.95	3.00	4.00	6.00	9.00	30.00
				Sample: 4 Yrs <= Horizon < 5 Yrs				
R^2	703,050	31.18	34.98	0.00	0.00	14.75	62.31	100.00
horizon	703,050	4.43	0.28	4.00	4.19	4.39	4.65	5.00
#Stocks	703,050	6.26	3.54	3.00	4.00	5.00	8.00	30.00

Table II: Forecasts informativeness: Trend by horizon

This table presents OLS estimates of time trend in analysts' forecasts' informativeness by sub-samples including observations in different annual forecasting horizon ranges. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Horizon (h) is the forecasting horizon measured as the number of days between the forecasting date and the date of actual earnings release divided by 365. Year Trend is a variable that takes the value of zero for the period 1983-1992 and increments by one every subsequent year divided by 25 so that the regression coefficient can be interpreted as the total increment in informativeness over the 1993-2017 period. We include fixed effects for industry (the main industry of analysts' portfolio), firms' size quintiles and age (based on average size and age in analysts' portfolio). In Panel A, the sample includes all analysts. In Panel B, the sample includes analysts issuing both short-term and long-term forecasts. Detailed variable definitions are provided in the Appendix. *t*-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)									
Sample:	h <1 Yr		1 Yr <= h <2 Yrs		2 Yrs <= h <3 Yrs		3 Yrs <= h <4 Yrs		4 Yrs <= h <5 Yrs	
OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: All analysts										
Year Trend	11.5***	11.0***	9.4***	8.4***	2.4	0.3	-11.5***	-7.2***	-20.0***	-13.9***
	(8.00)	(7.78)	(6.89)	(6.07)	(1.46)	(0.20)	(-5.12)	(-2.75)	(-5.42)	(-3.39)
Constant (83-92)	74.7***		55.0***		47.9***		44.3***		42.6***	
	(93.81)		(82.46)		(39.10)		(29.78)		(21.12)	
SIC2 FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Size FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Age FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	33,413,667	31,308,798	25,060,925	23,326,180	5,361,069	5,012,427	1,349,749	1,291,499	703,050	672,490
Panel B: Analysts making both short and long-term forecasts										
Year Trend	7.1***	5.9***	4.5**	2.1	-3.2*	-3.2*	-13.3***	-8.9***	-20.0***	-13.9***
	(4.13)	(3.72)	(2.32)	(1.05)	(-1.69)	(-1.70)	(-5.15)	(-2.98)	(-5.42)	(-3.39)
Constant (83-92)	78.4***		59.2***		50.2***		44.9***		42.6***	
	(72.41)		(50.62)		(40.56)		(27.27)		(21.12)	
SIC2 FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Size FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Age FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	4,411,947	4,217,939	3,815,445	3,639,981	2,040,510	1,960,221	1,195,965	1,151,999	703,050	672,490

Table III: Trend in the slope of the term-structure of forecasts informativeness

This table presents OLS estimates of time trend in the term-structure of analyst forecasts' informativeness (R^2). The dependent variable is the slope of the term-structure, measuring the change of forecasts' informativeness observed when horizon increases by one year. A negative slope indicates that forecasts' informativeness decreases with horizon. In column (1), the slope is calculated every year by regressing the average of R^2 by horizon on the horizon h (i.e., the number of days between the forecasting date and the date of actual earnings release divided by 365). In columns (2) and (3), the slope is calculated every year by 2-digit SIC industry by regressing the average of R^2 by horizon and industry on h. In columns (4) and (5), the slope is calculated every year by analyst by regressing the average of R^2 by horizon and analyst on h. Year Trend is a variable that takes the value of zero for the period 1983-1992 and increments by one every subsequent year divided by 25 so that the regression coefficient can directly be interpreted as the total change in slope over the 1993-2017 period. In Panel A, the sample starts in 1983. In Panel B, the sample starts in 1990. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by year. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dependent Variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS	(1)	(2)	(3)	(4)	(5)
Panel A: Whole sample					
Year Trend	-10.6***	-5.8***	-4.9***	-6.2***	-4.4**
	(-6.26)	(-5.50)	(-4.70)	(-7.38)	(-2.31)
Constant (83-92)	-6.6***	-10.0***		-10.0***	
	(-6.39)	(-20.05)		(-19.36)	
Analysts FE	-	-	-	No	Yes
N	32	775	769	3,826	3,725
Panel B: Excluding 80's					
Year Trend	-7.1***	-4.2***	-3.4***	-4.7***	-4.3**
	(-6.82)	(-3.92)	(-3.07)	(-7.51)	(-2.22)
Constant (90-92)	-8.6***	-11.0***		-11.0***	
	(-12.73)	(-20.01)		(-35.42)	
Analysts FE	-	-	-	No	Yes
N	25	686	681	3,694	3,583

Table IV: StockTwits' sample descriptive statistics

This table presents descriptive statistics for the main analyst-day-horizon variables in the Stocktwits' sample. R^2 measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. The forecasting horizon is measured as the number of days between the forecasting date and the date of actual earnings release divided by 365. #Stocks is the number of stocks covered by an analyst on a forecasting day. #Watchlist is the average number of users that have in their watchlist the firms covered by an analyst on a given day. #Messages is the average number of messages written about firms (in the last thirty days) that analyst covers on a given day. Auto-correlation is the average earnings' autocorrelation across the firms that an analyst covers on a given day. The other variables are control variable used in the analysis detailed in the Appendix. The sample covers the period from 2005 to 2017.

	N	Mean	St.Dev	Min	P25	P50	P75	Max
R^2	31,623,239	68.33	33.76	0.00	46.43	83.10	96.36	100.00
Horizon	31,623,239	1.26	0.93	0.00	0.54	1.11	1.77	5.00
#Stocks	31,623,239	10.37	5.46	3.00	6.00	9.00	13.00	30.00
#Watchlist	30,958,706	321	1,471	0	0	12	117	44,145
#Messages	30,958,706	11	41	0	0	2	8	1,304
Total Assets	29,390,791	11,738	32,854	0	1,548	4,616	12,635	2,087,821
Total Assets (Log)	29,390,791	8.35	1.54	-4.65	7.34	8.44	9.44	14.55
Age	29,392,408	22.97	12.41	1.00	13.43	20.24	29.90	68.00
Age (Log)	29,392,408	2.98	0.57	0.00	2.60	3.01	3.40	4.22
Cash Flow	29,383,877	0.05	0.12	-0.68	0.04	0.08	0.11	0.24
Cash	29,390,524	0.21	0.17	0.01	0.08	0.15	0.30	0.88
Debt	29,390,791	0.24	0.14	0.00	0.13	0.22	0.32	0.85
Q	29,366,118	2.29	1.05	0.71	1.54	2.00	2.74	7.34
Auto-correlation	29,364,398	0.67	0.21	-0.01	0.55	0.69	0.82	1.12

Table V: Social media data and forecasts informativeness by horizon

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data (from StockTwits). We consider different sub-samples including observations in different annual forecasting horizon ranges. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Social Media Data is an aggregate measure of analysts' exposure to StockTwits activity, measured first by firm and then averaged across the firms covered by analysts at time $t-1$. We normalize this variable by its standard deviation across analysts. In panel A, Social Media Data corresponds to the number of users that have the firm in their watchlist. The watchlist is set to zero when a firm is not covered/discussed on the platform. In Panel B, Social Media Data corresponds to the number of messages written about a firm from $t-30$ to $t-1$. The number of messages is set to zero when the stock is not covered/discussed on the platform. The sample period is 2005-2009 and both measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)							
Sample:	h <=1		1< h <=2		2< h <=3		h>=3	
OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Proxy for Social Media Exposure based on # Watchlist								
Social Media Data	1.62***	1.57***	1.18	0.51	-1.65***	-2.53***	-4.73***	-4.83***
	(3.90)	(4.03)	(1.07)	(0.47)	(-3.20)	(-4.78)	(-3.49)	(-3.20)
Analysts FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	14,026,800	13,006,543	11,502,199	10,612,608	3,929,446	3,648,151	1,500,165	1,438,756
Panel B: Proxy for Social Media Exposure based on # Messages								
Social Media Data	1.30***	1.03***	0.09	-0.68	-1.26	-2.04**	-3.34***	-3.00***
	(3.28)	(2.51)	(0.08)	(-0.63)	(-1.34)	(-2.03)	(-3.58)	(-3.02)
Analysts FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	14,026,800	13,006,543	11,502,199	10,612,608	3,929,446	3,648,151	1,500,165	1,438,756

Table VI: Social media data and forecasts informativeness by horizon: interaction approach

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data (from StockTwits). We consider all available analyst-day-horizon observations. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Social Media Data is an aggregate measure of analysts' exposure to StockTwits activity, measured first by firm and then averaged across the firms covered by analysts at time $t-1$. We normalize this variable by its standard deviation across analysts. Social Media Data corresponds to the number of users that have the firm in their watchlist, or number of messages written about a firm from $t-30$ to $t-1$. The forecasting horizon is measured as the number of days between t and the date of actual earnings release divided by 365, minus one so that the regression coefficient on the baseline variable Social Media Data can be interpreted as the unconditional effect on one-year informativeness. The sample period is 2005-2009 and both measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Social Media Data:		#Watchlist			#Messages	
Horizon × Social Media Data	-3.20***	-2.88***	-3.55***	-2.36***	-2.70***	-2.91***
	(-2.59)	(-3.06)	(-3.72)	(-3.88)	(-4.26)	(-4.31)
Social Media Data	0.47	-0.62	-1.29	0.47	-0.52	-1.19
	(0.50)	(-0.64)	(-1.29)	(0.62)	(-0.70)	(-1.61)
Horizon	-16.66***			-16.59***		
	(-33.86)			(-32.28)		
Analysts FE	Yes			Yes		
Date FE	Yes			Yes		
Analysts × Horizon FE		Yes	Yes		Yes	Yes
Date × Horizon FE		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,958,705	30,105,299	27,860,178	30,958,705	30,105,299	27,860,178

Table VII: Differential effects by social media users' investing horizon

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data (from StockTwits). We consider all available analyst-day-horizon observations. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Stocktwits' users self-declare their profile as investor, including their usual investing horizon, which they can define by declaring themselves as "Day Traders", "Swing Traders","Position Traders", or "Long-term investors". In columns (1) to (3), we proxy for Social Media Data using the number of messages written about the firm from $t-30$ to $t-1$ by users of each horizon category, which we average by analyst at time $t-1$, and then normalise by its standard deviation. Horizon is the forecasting horizon measured as the number of days between t and the date of actual earnings release divided by 365, minus one so that the regression coefficient on the baseline variable Social Media Data can be interpreted as the unconditional effect on one-year informativeness. The sample period is 2005-2009 and all measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)		
Social Media Data:		#Messages	
OLS	(1)	(2)	(3)
Horizon × (#Messages by Day Traders)	-1.06*	-0.87***	-0.97***
	(-1.86)	(-3.98)	(-3.83)
Horizon × (#Messages by Swing Traders)	-0.97	-0.88***	-0.87***
	(-1.46)	(-3.32)	(-3.56)
Horizon × (#Messages by Position Traders)	0.48	0.12	0.23
	(1.14)	(0.35)	(0.70)
Horizon × (#Messages by LT Investors)	-0.12	-0.02	-0.03
	(-0.22)	(-0.04)	(-0.05)
#Messages by Day Traders	0.84	0.50*	0.21
	(1.58)	(1.65)	(0.71)
#Messages by Swing Traders	-0.83	-0.94	-1.07
	(-1.17)	(-1.32)	(-1.59)
#Messages by Position Traders	0.92***	0.45	0.11
	(2.54)	(1.36)	(0.30)
#Messages by LTI Traders	0.04	-0.03	-0.06
	(0.11)	(-0.09)	(-0.17)
Horizon	-16.57***		
	(-31.87)		
Analysts FE	Yes		
Date FE	Yes		
Analysts × Horizon FE		Yes	Yes
Date × Horizon FE		Yes	Yes
Controls			Yes
N	30,958,705	30,105,299	27,860,178

Table VIII: Differential effects by analysts' processing constraints

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data (from StockTwits). We consider all available analyst-day-horizon observations. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Social Media Data is an aggregate measure of analysts' exposure to StockTwits activity, measured first by firm and then averaged across the firms covered by analysts at time $t-1$. We normalize this variable by its standard deviation across analysts. Social Media Data corresponds to the number of users that have the firm in their watchlist, or number of messages written about a firm from $t-30$ to $t-1$. The forecasting horizon is measured as the number of days between t and the date of actual earnings release divided by 365, minus one so that the regression coefficient on the baseline variable Social Media Data can be interpreted as the unconditional effect on one-year informativeness. #Stocks is the number of stocks covered by an analyst on a given forecasting day. The sample period is 2005-2009 and all measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Social Media Data:		#Watchlist			#Messages	
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Social Media Data × #Stocks	-0.50***	-0.23***	-0.23***	-0.29***	-0.14	-0.16*
	(-5.71)	(-3.38)	(-3.82)	(-5.14)	(-1.52)	(-1.89)
Horizon × Social Media Data	2.57	-0.13	-0.85	-0.16**	-0.09	-0.09
	(1.61)	(-0.10)	(-0.74)	(-2.42)	(-1.10)	(-1.16)
Horizon × #Stocks	-0.15***	-0.23***	-0.23***	-0.14***	-0.23***	-0.22***
	(-6.58)	(-8.67)	(-8.24)	(-5.87)	(-8.67)	(-8.36)
Social Media data × #Stocks	-0.32***	-0.19***	-0.16**	-0.16**	-0.09	-0.09
	(-3.34)	(-2.88)	(-2.26)	(-2.42)	(-1.10)	(-1.16)
#Stocks	-0.22***	-0.23***	-0.25***	-0.23***	-0.24***	-0.25***
	(-5.97)	(-6.95)	(-7.00)	(-5.98)	(-7.02)	(-6.93)
Social Media Exposure	4.09***	1.55	0.51	2.32***	0.47	-0.23
	(2.80)	(1.48)	(0.53)	(3.04)	(0.62)	(-0.32)
Horizon	-16.66***			-16.59***		
	(-33.86)			(-32.28)		
Analysts FE	Yes			Yes		
Date FE	Yes			Yes		
Analysts × Horizon FE		Yes	Yes		Yes	Yes
Date × Horizon FE		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,958,705	30,105,299	27,860,178	30,958,705	30,105,299	27,860,178

Table IX: Differential effects by earnings' auto-correlation

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data (from StockTwits). We consider all available analyst-day-horizon observations.The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Social Media Data is an aggregate measure of analysts' exposure to StockTwits activity, measured first by firm and then averaged across the firms covered by analysts at time $t-1$. We normalize this variable by its standard deviation across analysts. Social Media Data corresponds to the number of users that have the firm in their watchlist, or number of messages written about a firm from $t-30$ to $t-1$. The forecasting horizon is measured as the number of days between t and the date of actual earnings release divided by 365, minus one so that the regression coefficient on the baseline variable Social Media Data can be interpreted as the unconditional effect on one-year informativeness. Auto-correlation is the average earnings' autocorrelation in analysts' portfolios on a given day. The sample period is 2005-2009 and all measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in the Appendix. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Social Media Data:		#Watchlist			#Messages	
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Social Media Data x Auto-correlation	20.91***	11.49***	10.39***	8.00***	5.17**	4.70**
	(3.23)	(2.82)	(2.62)	(3.57)	(2.19)	(2.12)
Horizon × Social Media Data	-18.01***	-10.84***	-10.49***	-8.16***	-6.20***	-6.06***
	(-3.88)	(-4.34)	(-4.28)	(-5.54)	(-3.27)	(-3.38)
Horizon × Auto-correlation	11.04*	10.21***	9.80***	1.82*	0.53	0.71
	(1.95)	(3.07)	(3.12)	(1.96)	(0.72)	(0.95)
Social Media Exposure × Auto-correlation	20.91***	11.49***	10.39***	5.50**	5.20***	4.58**
	(3.23)	(2.82)	(2.62)	(2.16)	(2.67)	(2.35)
Auto-correlation	8.13***	8.39***	6.36***	8.18***	8.48***	6.46***
	(7.38)	(8.88)	(6.66)	(7.27)	(8.78)	(6.66)
Social Media Exposure	-7.49*	-7.95***	-8.22***	-3.65*	-4.14**	-4.38***
	(-1.80)	(-3.25)	(-3.33)	(-1.84)	(-2.45)	(-2.60)
Horizon	-18.07***			-17.99***		
	(-28.26)			(-26.91)		
Analysts FE	Yes			Yes		
Date FE	Yes			Yes		
Analysts x Horizon FE		Yes	Yes		Yes	Yes
Date x Horizon FE		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	28,711,790	27,865,669	27,840,732	28,711,790	27,865,669	27,840,732

A Appendix

Table A.1: Variable Definitions

Variable	Definition
All firm-level variables are converted into analyst-level variable by taking the average across all stocks the analyst covers	
#Messages	Number of StockTwits' messages posted about a given firm over the last thirty days (from $t-30$ to $t-1$).
#Stocks	Total number of distinct stocks covered by an analyst on a given day.
#Watchlist	Total Number of StrockTwits' users having a given firm in their watchlist.
Age	1+number of years in Compustat since inception.
Auto-correlation	Within firm quarterly net income (ibq item in Compustat) auto-correlation, obtained by regressing ibq over the lag of ibq over the last 2 years (without constant). We require that the regression has at least 4 observations.
Cash flow to assets	$ib + dp/at$ (from Compustat).
Cash to assets	che/at (from Compustat).
Debt to assets	$(dlc + dltt)/at$ (from Compustat).
Horizon	Number of days between the date at which the beliefs of the analysts are observed by the econometrician, and the date at which the actual earnings for the associated forecasted fiscal period are announced, divided by 365. When the earnings announcement date for the same forecasted fiscal period differs across firms covered by the analyst, we use the median date.
Tobin Q	$(at - ceq + chso * prcc_f)/at$ (from Compustat).
R^2	Informativeness of the forecasts made by an analyst on given day and for a given horizon. A higher R^2 indicates that the forecasts explain a larger fraction of the variation in realized earnings for the forecasted horizon, where the horizon corresponds to the number of days between a forecasting day and the date of actual earnings release divided by 365.

Table A.2: Social media data and analysts' forecasting activity

This table presents OLS estimates of the relationship between the propensity that an analyst issues a new forecast and social media activity. The sample covers the 2009-2017 period. The test is at the analyst-firm-day level. The dependent variable is a binary variable equals to one if the analyst issues a new forecast (or a revision) on a given firm during the day and zero if not. Social Media Data corresponds to the number of StockTwits' messages written about a firm during the prior thirty days. The number of messages is set to zero when the stock is not covered/discussed on the platform. Trading Volume is the total volume of trading on the firm during the prior thirty days. In Column (3), we impose that no news (from Capital IQ Key development dataset) is released about the firm during the day (otherwise the observation is removed from the sample). In Column (4), we impose that no news is released about the firm during the prior thirty days (otherwise the observation is removed from the sample). Detailed variable definitions are provided in Appendix. t-statistics in parentheses are based on standard errors clustered by firms. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Binary Variable (New Forecast=1)			
Social Media Data:	#Messages			
OLS	(1)	(2)	(3)	(4)
Social Media Data	0.0005***	0.0008***	0.0014***	0.0015***
	(2.97)	(4.29)	(8.82)	(2.70)
Trading Volume		-0.0011***	-0.0004***	0.0007*
		(-9.74)	(-4.12)	(1.86)
Analyst × Stock FE	Yes	Yes	Yes	Yes
Analyst × Date FE	Yes	Yes	Yes	Yes
Sample with no news at t	No	No	Yes	No
Sample with no news from t-30 to t	No	No	No	Yes
N	80,434,931	80,379,362	69,414,958	3,147,979