Does Big Data Improve Financial Forecasting? The Horizon Effect

Olivier Dessaint, Thierry Foucault, and Laurent Frésard*

December 4, 2020

ABSTRACT

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

Key words: Big data, Data Abundance, Financial analysts, 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 Copenhagen Business School, the Corporate Finance Webinar, INSEAD, the Universita della Svizzera Italiana, and the University of Geneva for useful comments. All errors are the authors' alone.

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 precision of investors' forecasts at various horizons are unknown. Yet, understanding these effects is important because many financial decisions rely on forecasts over multiple 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 precision 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. If data abundance impairs their long-term forecasts, it might negatively affect the informativeness of asset prices 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, credit card or internet traffic data about a given firm (e.g., satellite

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

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 (still) requires human judgment and methods of information processing that cannot be easily automated (e.g., 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 precision relatively more than the cost of producing long-term forecasts of the same precision. 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 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 forecasting short-term and long-term earnings are (related but) distinct tasks. The former requires primarily information on firm's assets in place, while the latter necessitates information on growth options. Thus, if the analyst puts more effort in sharpening the precision of a signal at a given horizon (e.g., by collecting and

[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.

processing more short-term information), the cost of increasing further the precision of the other signal increases as well.[3]

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 processing 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 precision of the analyst's signal about the innovation in the long-term earnings more than offsets the improvement in the precision 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 (i.e., when the cost of switching tasks is high).

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 different horizons. Specifically, we measure the overall informativeness of the forecasts of an analyst on a given forecasting day 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]

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

We implement this approach using the earnings' forecasts from I/B/E/S made by 14,379 analysts on 13,379 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

the fact that she invests little in information processing 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, 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.

has increased the most in our sample.[5]

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] Two aspects of StockTwits make it an appealing laboratory to precisely test the model's predictions. First, social media data like those from StockTwits mainly provide information about short-term prospects. Existing research indicates 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, the majority of StockTwits users in our sample self-identify as having short-term horizons. Second, analysts are likely to access and use social media data from StockTwits as a source of information. Indeed, data vendors such as Bloomberg or Thomson Reuters have gradually integrated StockTwits feed on their terminals for market professionals. We also report that analysts are more likely to make a new forecast about a stock following an increase in information produced on StockTwits about this stock.[7]

We measure the volume of data available on social media for a given stock by the

[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 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]This result holds even after controlling for trading activity and the flow of public news about a stock.

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 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 steepening of the informativeness term-structure should be more pronounced when social media data contains more short-term information (so that the marginal cost of producing short-term information drops more). Consistent with this prediction, the steepening of the informativeness term-structure 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 cost of switching forecasting tasks to increase with the number of stocks followed by an analyst. If this is the case, the model implies that the steepening of the informativeness term-structure 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.

In addition to supporting the model's predictions, these ancillary results also lessen potential concerns that the steepening of the term-structure of forecasts' informativeness might not be due to analysts' use of alternative social media data, but to unobserved variables correlated with social media activity (e.g., news arrival or firms' disclosure). Indeed, any candidate alternative explanation should not only explain the association

[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.

between analysts' exposure to social media data on the steepening of the informativeness of their forecasts over horizons, but also its cross-sectional variation (across key model's parameters). We believe that this strong requirement significantly reduces the scope for plausible alternative explanations.

II Related Literature

Our results add to the growing research studying the effects of progress in information technology and data abundance on financial markets. Existing theories on this issue (e.g., Abis (2018), Begeneau, Farboodi, and Veldkamp (2018), Dugast and Foucault (2018) or Farboodi and Veldkamp (2020)) posit that this evolution reduces the cost of accessing and processing information (or relaxes information capacity constraints) and focus on the implications for the informativeness of asset prices, the growth rates of small and large firms, or information acquisition choices by asset managers (e.g., Abis (2018)).

Correspondingly, a growing empirical literature analyzes how reductions to the cost of accessing and producing information due the digitization of data or the emergence of alternative sources of data (such as satellite images, geolocation data or social medias) affect financial markets and firms' decisions. For instance, Zhu (2019) and Grennan and Michaely (2019) find that the introduction of alternative data has a positive effect on proxies for stock price informativeness, while Katona, Painter, Patatoukas, and Zeng (2019) find no effect of the availability of satellite imagery for investors on price efficiency. Gao and Huang (2020) find the digitization of firms' regulatory filings (e.g., forms 10-Ks) and remote access to these filings (via the SEC EDGAR system) is associated with an increase in the informativeness of individual investors' order flow, the number of analysts covering a firm, and the precision of analysts' short-term forecasts.[9] 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 report 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 by investors, reducing the informational content of stock prices that

[9]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.

is new to firms' managers.

Our analysis differs in two important ways. First, we study how the availability of abundant data affects incentives to process information relevant for forecasting 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 relevant since mot financial decisions (e.g., asset valuations, portfolio allocations, or capital budgeting) require making forecasts about fundamental outcomes (e.g., cash-flows) that will occur at different dates in the future. Second, we do not focus on the informativeness of asset prices but on that of analysts' forecasts. This is important because analysts are important information providers and their recommendations or forecasts are informative and affect financial markets (see, for instance, Womack (1996) or Crane and Crotty (2020)). For this reason, our findings also add to the literature studying how progress in information technologies and data abundance affect the organization and output of security analysts (e.g., Grennan and Michaely (2020) or van Binsbergen, Han, and Lopez-Lira (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. There is one firm with two cash-flows (earnings), θ_{st} and θ_{lt}, realized at dates 2 (the short-term) and 3 (the long-term), respectively. At date 1, an analyst covering this firm announces her forecasts for its short-term and long term-earnings. 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. Thus, short-term and long-term earnings are correlated and this correlation increases with β.

The component of long-term earnings unrelated to short-term earnings represents, for instance, outcomes of R&D investments that cannot be predicted with information about short-term earnings (e.g., growth options).

[Insert Figure I about here]

Let f_{st} and f_{lt} 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:

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

where $\omega > 0$ and $\gamma \in [0.5, 1]$. One can interpret W as the total analyst's compensation from dates 2 to 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 $\gamma = 1/2$, both errors matter equally for her payoff. 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 precise long-term forecasts), her career concerns (the analyst's overall reputation should increase with the quality of her short-term and long-term forecasts) and discount rates.[10]

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

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

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, analysts meetings, industry reports, regulatory filings, news and scientific articles, social media, etc.) that possibly contain both short-term information (about θ_{st}) and long-term information (about e_{lt}).

[10]Results are identical if the analyst is paid at date 2 based on the realization of her forecasting error at this date (i.e., $(f_{st} - \theta_{st})^2$) and then at date 3 based on the realization of her forecasting error at this date ($(f_{lt} - \theta_{lt})^2$). 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.

After processing all data available to her, the analyst obtains two signals: (i) 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. Thus, $\Omega_1 = \{s_{st}, s_{lt}\}$. We assume that:

$$\begin{aligned} s_{st} &= \theta_{st} + \eta_{st} + \varepsilon_{st}, \\ s_{lt} &= e_{lt} + \eta_{lt} + \varepsilon_{lt}, \end{aligned} \tag{4}$$

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 firm's earnings, θ_{st} and θ_{lt}).

The analyst can reduce the noise coming from the $\varepsilon_j s$ in her signals by collecting short-term and long-term information. To formalize this idea, we assume that $\epsilon_j \sim \mathcal{N}(0, (Z - z_j)\xi_j^2)$, where z_j is the effort exerted by the analyst to increase the precision, $\tau_j(z_j)$, of her signal at horizon j, where $j \in \{lt, st\}$. In contrast, the analyst cannot learn about the noise coming from the ηs and we assume that $\eta_{jt} \sim \mathcal{N}(0, \kappa_{jt}^2)$ for $j \in \{lt, st\}$.

Thus, the precision of the analyst's signal about earnings at horizon $j \in \{st, lt\}$ is $\tau_j(z_j) = (\kappa_j^2 + (Z - z_j)\xi_j^2)^{-1}$.[11] The larger is the analyst's effort to collect information about short-term earnings, z_{st}, the higher is the precision of s_{st}, her signal about these earnings. If the analyst chooses the largest possible effort for the production of this signal ($z_{st} = Z$), she obtains a signal of precision $1/\kappa_{st}^2$. Thus, parameter κ_{st} controls the highest precision that the analyst can achieve for her short-term signal. It measures the extent to which relevant information about short-term earnings is available in the data. If there is a lot relevant information, $1/\kappa_{st}^2$ is high and the analyst can, with sufficient effort, produce a signal of high quality about short-term earnings. Parameter ξ_{st} controls both the precision of the analyst's short-term signal in the absence of effort and the marginal benefit of the analyst's effort. Indeed, the higher is ξ_{jt}^2, the smaller is the precision of the analyst's short-term signal in the absence of effort ($\tau_{st}(0)$) and the higher is the increase in the precision of this signal for a one unit increase in effort. The interpretation of parameters κ_{lt} and ξ_{lt} in the specification for the long-term signal, s_{lt}, are identical.[12]

[11] The effort of the analyst to acquire information for a specific horizon is specific to this horizon. This is a natural assumption: Data collected about the unique component of long-term earnings cannot be used, by definition, for forecasting short-term earnings.

[12] See Myatts and Wallace (2012) for a similar information structure in a different context.

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

$$\text{Max}_{f_{st},f_{lt}} \bar{W}(f_{st}, f_{lt}; s_{st}, s_{lt}), \tag{5}$$

where $\bar{W}(f_{lt}, f_{st}; s_{lt}, s_{st})$ is given in eq.(3). 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:[13]

$$\begin{aligned} f_{st}^* &= \mathsf{E}(\theta_{st} \,|s_{st}) \,, \\ f_{lt}^* &= \mathsf{E}(\theta_{lt} \,|s_{st}, s_{lt}) \,. \end{aligned} \tag{6}$$

To simplify the analysis, it is convenient to assume that the analyst has improper priors about θ_{st} and e_{st}.[14] In this case, we have:

$$\begin{aligned} f_{st}^* &= s_{st}, \\ f_{lt}^* &= s_{st} + s_{lt}. \end{aligned} \tag{7}$$

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

$$C(z_{st}, z_{lt}) = az_{st}^2 + bz_{lt}^2 + cz_{st}z_{lt}, \tag{8}$$

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 processing a signal increases in its precision then it should naturally increase in the precision achieved for other signals as well. This specification captures the idea that forecasting short-term and long-term earnings are genuinely different tasks. The former requires understanding the firm's assets in place, whereas the latter requires knowledge of the firm's growth potential. If the 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

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

[14]This means that the prior variances of these variables are infinitely large. This assumption is not key but simplifies expressions in many places.

effort of collecting additional information, be it short-term ($a > 0$) or long-term ($c > 0$). One can also interpret c as capturing switching costs associated with multitasking: If the analyst devotes much time to the cost of forecasting long-term earnings then switching to the task of forecasting short-term earnings is costly and vice versa.[15] For reasons that will become clear below, we assume that $4ab > c^2$.[16]

The analyst chooses her efforts at date 0 (the search period), i.e., before obtaining her signals and formulating her forecast to maximize her ex-ante expected payoff net of information processing costs. That is, z_{st} and z_{lt} are chosen to solve:

$$\text{Max}_{z_{st},z_{lt}} J(z_{st}, z_{lt}) = \mathsf{E}(\bar{W}(f^*_{st}, f^*_{lt}; s_{st}, s_{lt})) - C(z_{st}, z_{lt}), \tag{9}$$

where the analyst's forecasts at date 1, f^*_{st} and f^*_{lt}, are given by eq.(7) (i.e., are chosen optimally). We next analyze the solution to this problem and its implication for the informativeness of analysts' forecasts.

B Optimal Information Processing 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:

$$\begin{aligned} J(z_{st}, z_{lt}) &= \omega - \gamma\, \mathsf{E}((f^*_{st} - \theta_{st})^2) - (1-\gamma)\, \mathsf{E}((f^*_{lt} - \theta_{lt})^2) - C(z_{st}, z_{lt}), \\ &= \omega - \gamma\, \mathsf{E}(\mathsf{Var}(\theta_{st} \,|s_{st})) - (1-\gamma)\, \mathsf{E}(\mathsf{Var}(\theta_{lt} \,|s_{lt}, s_{st})) - C(z_{st}, z_{lt}), \\ &= \omega - (\gamma + (1-\gamma)\beta^2))\, \mathsf{Var}(\theta_{st} \,|s_{st}) - (1-\gamma)\, \mathsf{Var}(e_{lt} \,|s_{lt}, s_{st}) - C(z_{st}, z_{lt}), \end{aligned} \tag{10}$$

where the last line follows from the fact that (i) $\mathsf{Var}(\theta_{jt} \,|s_{jt}))$ does not depend on the realization of s_{jt} because θ_{jt} and s_{jt} are normally distributed, and (ii) 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

[15] Switching costs associated with multitasking are well documented in the psychological literature. See, for instance, Monsell (2003).

[16] Given the cost function specified in eq. (8), one may wonder why analysts do not specialize in only one task. Under the current practice of equity research, the objective of the analyst is to evaluate the market price of a stock to make a buy or sell recommendation. This evaluation requires making forecasts about short-term and long-term earnings. Fixed costs associated with coverage initiation make it difficult to separate those two tasks. Besides, as in the model, forecasting the long-term earnings depends on the short-term forecast, which renders specialization unlikely.

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:

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

Let $h(\beta,\gamma) \equiv (\gamma + (1-\gamma)\beta^2)$. Writing the first-order conditions of the analyst's problem at date 0 (eq.(10)), we deduce that the analyst's optimal efforts in information processing, z_{st}^* and z_{lt}^*, are :

$$\begin{aligned} z_{st}^* &= \text{Min}\{\frac{2bh(\beta,\gamma)\xi_{st}^2 - c(1-\gamma)\xi_{lt}^2}{4ab - c^2}, Z\} \\ z_{lt}^* &= \text{Min}\{\frac{2a(1-\gamma)\xi_{lt}^2 - ch(\beta,\gamma)\xi_{st}^2}{4ab - c^2}, Z\}. \end{aligned} \tag{11}$$

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 $\frac{ch(\beta,\gamma)}{2a(1-\gamma)} < \frac{\xi_{lt}^2}{\xi_{st}^2} < \frac{2bh(\beta,\gamma)}{c(1-\gamma)}$, the solution is interior in the sense that $0 < z_j^* < Z$, 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.(11) 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:

$$\frac{\xi_{lt}^2}{\xi_{st}^2} < \frac{h(\beta,\gamma)(c+2a)}{(1-\gamma)(c+2b)}. \tag{12}$$

Suppose that $\xi_{lt}^2 = \xi_{st}^2$, $a = b$ and $\gamma = 1/2$. In this case, the marginal cost and benefit of processing 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 the common component of short-term and long-term earnings is sufficiently large relative to the unique component of long-term earnings (i.e., β is large enough). The reason 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 analyst's earnings forecast at given horizon is more informative if it enables an external observer to reduce his uncertainty about the firm's earnings at this horizon by a greater amount. 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 realized earnings at this horizon conditional on the analyst's forecast at this horizon.[17] That is:

$$I_j \equiv \mathsf{Var}(\theta_j \mid f_j^*)^{-1} \quad for \quad j \in \{st, lt\} \tag{13}$$

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:

$$I_{st} = \mathsf{Var}(\theta_j \mid s_{st})^{-1} = (\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2)^{-1}, \tag{14}$$

where for the second equality we use the fact that the analyst has diffuse priors. Moreover, as $f_{lt}^* = s_{lt} + s_{st}$, we have:

$$I_{lt} = \mathsf{Var}(\theta_{lt} \,|f_{lt}^*)^{-1} = (\beta^2(\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2) + \kappa_{lt}^2 + (Z - z_{lt}^*)\xi_{lt}^2)^{-1}. \tag{15}$$

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 $\beta > 0$.[18]

As explained in the introduction, our hypothesis is that alternative data (e.g., satellite

[17] This is similar to the definition of price informativeness in rational expectations models. See for instance Grossman and Stiglitz (1980).

[18] Instead of considering the informativeness of each forecast separately at a given horizon, one could also consider the *joint* informativeness of both analysts' forecasts for short-term and long-term earnings, i.e., $I_{jt}^{joint} = \mathsf{Var}(\theta_{jt} \,|f_{st}^*, f_{lt}^*)^{-1})$ for $j \in \{st, lt\}$. Note that eq.(7) implies that observing the short-term and long-term analyst's forecasts is informationally equivalent to directly observing the analyst's signal, s_{st}, s_{lt}. As s_{lt} does not contain information about short-term earnings and $f_{st}^* = s_{st}$, the informativeness of the joint analyst's forecasts for short-term earnings is the same as the informativeness of he short-term forecast: $I_{st}^{joint} = I_{st}$. Moreover, under our assumption that the analyst has diffuse priors, one can also show that the informativeness of the analyst's long-term forecast for long-term earnings is the same as the informativeness of her joint forecasts for long-term earnings: $I_{lt}^{joint} = I_{lt}$.

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. Hence, to study the effect of this evolution, we analyze how a change in a, the parameter that determines the rate at which the marginal cost of producing the short-term signal increases with its precision, affects the analyst's choice of her efforts to produce short-term and long-term information (z^*_{st} and z^*_{lt}), holding other determinants of the total cost of processing information (b and c) constant.

Using eq.(11), we obtain (when the solution to the analyst's problem at date 0 is interior):

$$\begin{aligned} \frac{\partial z^*_{st}}{\partial a} &= -\frac{4b}{(4ab-c^2)} z^*_{st} < 0, \\ \frac{\partial z^*_{lt}}{\partial a} &= \frac{2c}{(4ab-c^2)} z^*_{st} > 0. \end{aligned} \tag{16}$$

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 precision of the short-term signal (eq.(16) shows that z^*_{st} increases when a decreases). Thus, the informativeness of her short-term forecast unambiguously increases with data abundance because:

$$\frac{\partial I_{st}}{\partial a} = (\frac{\partial z^*_{st}}{\partial a}) \frac{\xi^2_{st}}{(\kappa^2_{st} + (Z - z^*_{st})\xi^2_{st})^2} < 0. \tag{17}$$

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. With the advent of social medias, they can now, at a low cost, obtain opinions about a firm's prospects from a large pool of investors and use this information as an input for forecasting future earnings (in addition to other, more traditional, sources of information).[19] Even

[19]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

though the cost of accessing 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 cognitively more demanding for the analyst to focus on the firm's long-term prospects after having spent 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.(16), a drop in the marginal cost of obtaining short-term information leads the analyst to reduce her effort to collect long-term information ($\frac{\partial z^*_{lt}}{\partial a} > 0$ iff $c > 0$).

These effects have an ambiguous impact on the informativeness of the analyst' forecast 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. The second effect dominates if and only if $\beta^2 < \frac{c}{2b}\frac{\xi^2_{lt}}{\xi^2_{st}}$. Indeed, using eq.(15), we obtain that:

$$\frac{\partial I_{lt}}{\partial a} = (\beta^2\xi^2_{st}\frac{\partial z^*_{st}}{\partial a} + \frac{\partial z^*_{lt}}{\partial a}\xi^2_{lt})I^2_{lt} = -(\frac{2(2\beta^2\xi^2_{st}b - c\xi^2_{lt})}{(4ab - c^2)})z^*_{st}I^2_{lt}. \tag{18}$$

Thus, when $\beta^2 < \frac{c}{2b}\frac{\xi^2_{lt}}{\xi^2_{st}}$, 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. This is always the case for (i) firms with low earnings' autocorrelation (β low enough) and (ii) analysts with a high cost of multitasking (c high enough).*

In the rest of the paper, we test these predictions. In Section IV, we first use a large

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.

panel of analysts' earnings forecasts over multiple horizons combined with actual earnings' realizations to measure the informativeness of analysts' forecasts at various horizons. In this way, we characterize the term-structure of each analyst' forecasts informativeness. Then in Section V, we study the long-run evolution of this term structure, conjecturing that data abundance has increased over time. Finally, in Section VI, we present our main test. Namely, we exploit cross-sectional and time variation in analysts' exposure to social media data to generate changes in the volume of data to which they have access and test whether an increase in this volume distorts the term-structure of their forecasts informativeness.

Remark: Existing research indicates 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:

$$f_j^{r*} = f_j^* + B_j. \tag{19}$$

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.

IV Data and Measurements

A Earnings Forecasts and Realizations

We construct a large sample of forecasts' informativeness at different horizons using analysts' forecasts of earnings per share (EPS) and net income (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.[20] We eliminate forecasts with missing announcement dates, analyst code, or broker code. When a given analyst issues 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 for 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 forecasts to build our main measure of "earnings" forecast. If an analyst issues 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 issues only an EPS forecast, we convert it into a net income forecast 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, in a way consistent with I/B/E/S's adjustments for these splits.

Next, we match earnings forecasts to realized earnings reported in the I/B/E/S Actual File. 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, otherwise, the number of shares from CRSP. Then, to build our final sample of earnings forecasts, we apply the following criteria. First, all earnings forecasts must be about a fiscal year ending between 1983 to 2017. Second, we require that actual earnings for the forecasted fiscal period and total assets from Compustat at the end of the forecasted fiscal period are not 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 disregard earnings forecasts that are in absolute value ten times greater than the firm's total assets at the end of the forecasted fiscal period.[21]

[20] 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.

[21] For the same reason, we also impose that actual net income (in absolute value) is not greater than total assets at the end of the forecasted fiscal period.

B Measuring Forecasts Informativeness

We construct a daily measure of informativeness by analyst and forecasting horizon.[22] 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. 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 their earnings.[23] 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.[24] Since analysts do not update their forecast daily, the panel has gaps, which we fill using the last available forecast whenever it is possible. To avoid stale forecasts, we only consider the last available forecast if it is not older than one year. At the end of this process, a given analyst-day-horizon assembles a collection of forecasts issued by the analyst about various firms for a given forecasting horizon. We provide an illustrative example of this process in Appendix A.

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 of the following regression:

$$e_j = k_0 + k_1 \hat{e}_j + \nu_j, \tag{20}$$

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,

[22] We build a (high-frequency) daily measure to be able to fully exploit the granularity of our data in the tests in which we vary analysts' exposure to data abundance in Section VI.

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

[24] 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.

respectively.[25] By definition, the R^2 of this regression is:

$$R^2_{i,t,h} = 1 - \frac{\mathsf{Var}(\nu_j)}{\mathsf{Var}(e_j)} = 1 - \frac{\mathsf{Var}(e_j|\widehat{e_j})}{\mathsf{Var}(e_j)}. \tag{21}$$

A higher $R^2_{i,t,h}$ means that analyst i's forecast of earnings at horizon h on day t explains a larger fraction of the (cross-sectional) variation in realized earnings at date $t+h$ for firms covered by the analyst. In this sense, a higher $R^2_{i,t,h}$ means that analyst i's forecasts at horizon h are more informative. Thus, we use $R^2_{i,t,h}$ as our measure of forecast informativeness at horizon ^{h} for each analyst in our sample.

This measure is closely related to our theoretical measure of price informativeness (given in eq.(13)), i.e., the inverse of the variance of an asset cash-flow at a given horizon conditional on the analyst's forecast of this cash-flow. Indeed, $\mathsf{Var}(e_j|\widehat{e_j})$ is a proxy for this variance and $R^2_{i,t,h}$ is inversely related to $\mathsf{Var}(e_j|\widehat{e_j})$ normalized by the (cross-sectional) variance of firms' earnings t horizon h. This normalization enables us to control for "prior" (before information acquisition) uncertainty about future earnings for the portfolio of firms covered by a specific analyst. This is important because a reduction of prior uncertainty also reduces $\mathsf{Var}(e_j|\widehat{e_j})$. By normalizing $\mathsf{Var}(e_j|\widehat{e_j})$ by $\mathsf{Var}(e_j)$, we neutralize this mechanical source of variation in $\mathsf{Var}(e_j|\widehat{e_j})$ and can therefore better attribute changes in the informativeness of the analyst's forecasts at a given horizon to variations in other sources of variations in this informativeness (e.g., the analyst's effort for producing information at this horizon).

Note that $R^2_{i,t,h}$ is analyst's specific, not analyst and firm specific as in the model. In effect, we are treating each pair $(e_j, \hat{e}_j)$ for a given analyst at a given horizon as different realizations of the analyst' forecast of the firm's earnings at a given horizon and the realization of this earnings (the pair $(\theta_j, f^j{*_j})$ for $j \in \{st, lt\}$) in our model.[26] The implicit assumption is that the distribution of firms' normalized earnings at various horizons are similar across firms in a given analyst's portfolio.[27]

[25] 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 financial statements on day t. One drawback of this alternative approach is that our measure of informativeness of an analyst's forecasts at a given horizon can change even when analysts do not update their forecasts (because the normalization changes).

[26] 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).

[27] As analysts tend to follow firms with similar product market characteristics, heterogeneity across

We obtain $R^2_{i,t,h}$ by estimating regression (20) for each available analyst-day-horizon collection. Of course, $R^2_{i,t,h}$ can be estimated only when we have at least three forecasts observations at horizon h by a given analyst on day t. When this is not the case, $R^2_{i,t,h}$ is non available. Moreover, we require (i) the number of observations for estimating eq.(20) to be less than thirty (to avoid using forecasts issued by teams rather than individual analysts), (ii) the horizon of the forecast to be greater than one day and smaller than five years. Finally, to limit the effect of outliers coming from lower power in some estimations with few observations, we drop estimates of $R^2_{i,t,h}$ when the estimated slope of eq.(20) (k_1) is in the first percentile in each tail of its distribution, and set $R^2_{i,t,h}$ to zero when the estimated k_1 is negative.

This procedure yields a sample containing 65,888,460 analyst-day-horizon observations of R^2, obtained from 14,379 distinct analysts who issued forecasts about 13,849 distinct firms with forecasting horizon ranging between one day and five years.

C The Term-Structure of Forecasts' Informativeness

Table I presents the summary statistics. 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, 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 disclose and revise their short-term forecasts more often than their long-term forecasts.[28] Second, in many instances we do not observe earnings' realizations associated with long-term forecasts because firms stay less than 5 years in the sample, or because they disappear before their earning is realized.[29]

firms in an analyst's portfolio is usually low (especially after normalizing by firms' size as we do).

[28]Note that the fact that, at a given date, an analyst only discloses a short-term forecast does not imply that she did not forecast long-term earnings at the date. Because both types of forecasts are needed to assess firms' valuation and make investment recommendations, the prevalence of short-term forecasts in I/B/E/S likely reflects analysts' reporting choice.

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

[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 forecasting horizons of less than one year, compared to about 1.3 million observations for forecasting horizons ranging between three and four years. The number of firms covered by an analyst also varies across forecasting horizons, with 8.14 firms covered for horizon less than one year compared to 6.70 for horizons ranging between three and four years.[30]

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. Thus, the term-structure of analysts' forecasts informativeness is downward-sloping. To better illustrate the shape of this term-structure, we regress $R^2_{i,t,h}$ on dummy variables capturing each (daily) horizon (from one day to five years). Figure II plots the estimated coefficients (together with their 90% confidence intervals) and confirms that forecasts at shorter horizons are significantly more informative than forecasts at longer horizons.

A visual inspection of Figure II (and Table I) suggests that the informativeness of an analyst's forecast decays quickly with the horizon of this forecast. Indeed, 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 for 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 deteriorates by about 12 percentage points for an annual increase in their forecasting horizon.

[30]Thus, our estimates of $R^2_{i,t,h}$ at longer horizons are less precise since they are obtained from estimations of eq.(20)) with fewer observations.

V Long Run Evolution

Clearly, the volume and diversity of available data has increased over time, in particular due to the digitization of vast amount of data. This process has accelerated in recent years but it started long ago in financial markets (e.g., the SEC requires firms to report regulatory files such as 10-Ks in electronic format since 1993). Our model predicts that, other things equal, this evolution should trigger a reallocation of analysts' efforts toward the production of short-term information, leading to an increase in the informativeness of their short-term forecasts and possibly a drop in the informativeness of their long-term forecasts. Thus, we first study the long run evolution of the term structure of analysts' forecasts informativeness, being fully cognizant that other factors than data abundance may explain this evolution.

Figure III displays the term-structure of analysts' forecasts informativeness for the periods 1983-2000 and 2001-2017. It suggests that this term-structure has indeed become steeper over the second part of our sample, with long-term (short-term) forecasts becoming markedly less (more) informative after 2000. To formally test whether this shift 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 for each forecasting horizon sub-sample. This counter is set to zero before 1992 and increases 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 (less than one year and two years) 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 (more than three and four years) has materially deteriorated, with coefficients on the trend of -11.5 and -20. 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 fixed effects control for changes in forecasts' informativeness that could stem from changes in the composition of 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 term-structure of analysts' forecasts informativeness, we estimate its slope as we did for Figure II but year by year starting in 1983. Figure IV shows how that slope has changed over time. The slope of the term-structure of analysts forecasts' informativeness becomes 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.

VI Social Media Data and Forecasts Informativeness

The informativeness of analysts' short-term forecasts has improved in the long run while the informativeness of long-term forecasts has decreased. Of course, as we already emphasized, there might be many other factors than data abundance explaining this evolution. For instance, an increased focus on short-term earnings by investors and corresponding changes in analysts' compensation schemes (an increase in γ in the model) could also generate such an evolution. In this section, to better isolate the role of data abundance on the term-structure analysts' forecasts informativeness, we exploit variation in the volume of data relevant about the short-term available to analysts that is plausibly unrelated to other factors affecting their incentives to allocate their effort between the production of short-term and long-term information.

Specifically, we use the introduction and expansion of StockTwits, a social networking platform for investors where users can publicly share their opinions about stocks and capital markets, as a proxy for a change in the volume of social media data available to analysts. Our premise is these opinions expand the possible sources of short-term information for analysts and thereby decreases their cost of increasing the precision of their forecasts about short-term earnings (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 to this platform can post messages of up to 140 characters and can use $cashtags with stocks' ticker symbols to link their messages to particular firms. Users of StockTwits and its services include, for instance, retail investors, finance professionals (e.g., analysts) and journalists. Several recent academic papers use data from StockTwits to address various questions (e.g., Cookson and Niessner (2020), Giannini, Irvine, and Shu (2019) or Cookson, Engelberg, and Mullins (2020)).

We obtained data from StockTwits for all messages posted between January 1, 2009 and Decembre 31, 2017. Similar to Cookson and Niessner (2020), for each message, we observe the user identifier, its date, its content, and the associated $cashtags with the corresponding tickers (a message can be associated with multiple tickers). We also observe specific information about both users and stocks. As for users, we have access to self-declared information provided when they registered on the platform, including their name and their investment horizon. Moreover, for each stock discussed on StockTwits, we know its listing venue and its "watchlist", i.e., the number of users who declare following 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]

Figure V shows the evolution of the number of users and their posting intensity on StockTwits. It shows that the intensity of activity on StockTwits has dramatically increased since its creation. 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 number of investors on a stock watchlist also increases sharply over time, up to about 2,000 in 2017. The lower panels of Figure V displays the evolution of the distributions of the daily numbers of messages and investors on a stock's watchlists from 2009 to 2017. We note a substantial and increasing heterogeneity in the availability of StockTwit messages across firms. Our tests exploit this time-series and cross-sectional variation in the availability of social media data.

B Analysts' Exposure to Social Media Data

We use two distinct stock-level daily measures of data abundance based on the recent activity of StockTwits' users. Specifically, we measure the amount of social media data on day t for stock i either by (i) the total number of investors in stock i's watchlist on day $t-1$, or (ii) the total number of messages "cashtagging" stock i 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 about a stock means that more information is available about this stock.

We then measure the amount of social media information available to an analyst on a given day – the analyst's *exposure* to social media – by the average amount of social media data available for the stocks in the analysts' portfolio on this day. If no social media data is available for a stock on a given day, the analyst's exposure to social media information is set to zero. Hence, an analyst is more exposed to social media data when the stocks she covers have more users in their watchlists or are more frequently discussed in StockTwits messages. For our tests, we consider all analyst-day-horizon observations in our sample (i.e., with available forecast informativeness estimates, R^2) between 2005 and 2017. The resulting sample (henceforth the "StockTwits sample") contains 30,958,705 observations.

[Insert Table IV about here]

Table IV presents summary statistics. The average forecast informativeness is equal to 68.33%, which is similar to our estimate for the whole sample. 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 in the whole sample). Importantly, our two measures of analysts' exposure social media data display large variability. The average number of users in the watchlists of firms covered by the average analyst is equal to 321 with a standard deviation of 1,471. Similarly, the average number of messages for firms covered by the average analyst is equal to 11 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 financial statements and aggregated at the analyst-day level (and detailed in the Appendix).

C Relevance Conditions

Our tests using Stocktwits data rely on two conditions. The first condition is that social media data like those from StockTwits mainly provide information about short-term prospects (e.g., earnings) of firms discussed on those media. The second condition is that analysts use these data as a source of information. In this section, we argue that these two 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 (see, for instance, Chen, De, Hu, and Hwang (2014), Jame, Johnston, Markov, and Wolfe (2016), Renault (2017), or Bartov, Faurel, and Mohanram (2020)). Interestingly, StockTwits' users can self-declare one of four investment horizon category: "day trader", "swing trader", "position trader", and "long-term investors".[31] Using this information, Figure VI displays the breakdown of messages by users' declared horizons. The vast majority of StockTwits' messages stem from users that are either "day traders" (35.4%) or "swing traders" (49%), which is consistent with our conjecture that social media data mainly provide short-term information. In contrast, only a small fraction of posts are issued by users declaring a long-term horizon, either "position traders" (6.2%) or "long-term investors" (8.6%).

[Insert Figure VI about here]

Second, several indicators suggest that analysts are indeed exposed and sensitive to the information contained in StockTwits' activity. Firstly, 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 use social media data, we report in the Appendix (see Table **??**) several analyses indicating that analysts are significantly more likely to issue (or revise) a forecast on a given firm and day following an increase in activity of 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), we find that 35% (of 7,656 distinct analysts) of analysts' names exactly match that of

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

active StockTwits' users (i.e., users that have posted at last one message). Arguably, the matching between I/B/E/S and StockTwits is imperfect. However, although an account is not required to follow messages on StockTwits, this finding suggest that some analysts indeed possess StockTwits' accounts and are therefore following information generated on this social media.

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 estimate the following baseline specification:

$$R^2_{i,t,h} = \lambda(\text{Social Media Data})_{i,t-1} + \Gamma\text{Controls}_{i,t-1} + \eta_i + \eta_t + \omega_{i,t,h}, \tag{22}$$

where $R^2_{i,t,h}$ is the informativeness of analyst i's forecasts available at time t for the forecasting horizon h, and "*Social Media Data*" is analyst i's exposure to social media data at time $t-1$, measured by either the average number of users in the watchlists of stocks covered by the analysts or the number of past messages on StockTwits about these stocks. The baseline specification includes analysts fixed effects to absorb any time-invariant differences across analysts (e.g., their genuine forecasting ability) and time 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.[32] We cluster the standard errors of $\omega_{i,t,h}$ by forecasted fiscal period. To measure how analysts' forecasts informativeness changes after the introduction of StockTwits, the sample starts in 2005, i.e., five years prior to StockTwits' foundation.

The coefficient of interest in eq.(22) 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

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

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 eq.(22) separately across four distinct groups of horizons (with and without controls), ranging from one year or less ($h \leq 1$) to more than three years ($h \geq 3$).[33]

[Insert Table V about here]

Table V presents the results. In Panel A, social media data is measured by the average number of users in the watchlists of stocks covered by the analyst and in Panel B, it is measured by the average number of prior messages about these stocks. To facilitate economic interpretation, in either case, we standardize the variable "Social Media Data" by its sample standard deviation. 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 significantly reduces 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 ($h \geq 3$).

Across both panels, a one standard deviation increase in analysts' exposure to social media data leads to a drop in the informativeness of long-term forecasts of about 1.48% to 1.64%, and an improvement in the informativeness of their short-term forecast of about 0.37% and 0.54%. In relative terms, the estimated decline of long-term forecasts' informativeness is about three times larger than the corresponding improvement in the short-term (e.g., compare coefficients of -1.48 and -1.55 in columns (8) to 0.37 and 0.53 in columns (2)).[34]

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

[34]From a different perspective, a one standard deviation increase in analysts' exposure to social media data leads to a drop in the informativeness of long-term forecasts amounting to 4.3% (4.7%) of its sample standard deviation, and to an increase in the informativeness of short-term forecasts amounting to 2.2% (1.9%) of its sample standard deviation.

[Insert Table VI about here]

To provide a different perspective on the economic magnitude of these effects, we modify the baseline eq.(22) by pooling together analyst-day-horizon observations across all horizons, 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).[35] 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. Thus, greater exposure to social media data makes the term-structure of informativeness steeper. Column (1) (respectively column (2)) indicates that, for a given increase in social media data, the informativeness of analysts' forecasts decreases more than in the absence of such exposure (the baseline). Specifically, an annual increase of the forecasting horizon (e.g., from $h = 1$ to $h = 2$) reduces the informativeness of an analyst' forecast by 16.66% (16.59%) in the absence of social media exposure and by 0.86% (0.77%) more for a one standard deviation increase in social media exposure (a drop at a rate of about 5% per year).

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.

We report two additional robustness tests in the Appendix. First, we show in Table A.3 that our main result is unlikely due to analysts' changing their coverage in response to increased social media data (e.g., intiate coverage of firms with less social media data). In particular, we show that our conclusion holds in a subsample of analysts with "stable" portfolios, defined as those displaying a similarity in their portfolio between t and $t-1$ greater than 90%. Second, we report in Table A.4 that our results hold when we control for trading volume (averaged across the stocks followed by each analyst), that could arguably correlate with both social media activity and analysts' informativeness across

[35] More specifically, we estimate: $R^2_{i,t,h} = \lambda(\text{Social Media Data}) \times (h-1)_{i,t-1} + \varphi(h-1) + \kappa(\text{Social Media Data}) + \Gamma\text{Controls}_{i,t-1} + \eta_i + \eta_t + \omega_{i,t,h}$.

horizons (e.g., firms' disclosing material information affecting analysts' forecasts across horizons).

E Additional Predictions and Ancillary Results

To further document the economic channel at play in the model, we test three unique ancillary predictions of our theory. Indeed, the steepening of the term-structure of analysts' forecasts informativeness should be more pronounced (i) when social media data contains more short-term information (i.e. when the marginal cost of producing short-term information a decreases), (ii) when the cost assiocated with switching tasks is high (i.e. when c is more negative), and (iii) when firms' earnings are less auto-correlated (i.e. when β is low). We find broad support for these predictions.

E.1 Users' Investing Horizon (a)

Our premise is that social media like StockTwits mainly contain short-term information. Thus, coverage of stocks by social media reduces the cost of processing short-term information for analysts relative to the costs of processing long-term information. This effect should be stronger for stocks that attract relatively more users with short-term horizons. This logic implies that the positive (negative) association between analysts' exposure to social media data and the informativeness of their short-term (long-term) forecasts should be stronger for stocks that are followed by a greater number of short-term users on StockTwits.

To test this precition, we exploit the heterogeneity in investing horizon across StockTwits' users. We posit that users who define themselves as "day traders" are more likely to collepcively roduce 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 ($t-1$) by each category of trader (i.e., "day trader", "swing trader", "position trader" and "long-term investor") for each stock covered by an analyst and compute the average number of messages for each category across stocks covered by the analyst. We then reestimate the specifications reported in Table VI (Columns (4) to (6)) breaking down the average number of messages for each category. That is, we measure the effect of analysts' exposure to messages in each category rather than aggregating all the categories together.

[Insert Table VII about here]

Table VII shows that the negative (positive) reltion between analysts' exposure to social media data on the informativeness of their long-term (short-term) forecasts is significantly stronger for stocks discussed by users with short-term horizons. In fact, the interaction terms between the average number of messages for the stocks covered by analyst and her forecasting horizon is significantly negative only for messages written by "day traders" (coefficients ranging between -0.87 and -1.06) and "swing traders" (coefficients ranging between -0.88 and -0.97). For other categories of users ('position traders" and "long-term investors"), there is no significant relationship between analysts' forecasts informativeness and the number of messages about the stocks they cover on StockTwits. Overall, findings in Table VII support our hypothesis that the steepening of the term-structure of forecasts informativeness stems from the preponderance of short-term information in social media data.

E.2 Cost of Switching Tasks (c)

Forecasting firms' long-term earnings (e.g., coming from growth options) is a task distinct than forecasting their short-term earnings (e.g., coming from assets in place) and the marginal cost of effort (or attention) for the first task increases with the effort allocated to the second task (and vice versa). In our model, this "multi-tasking cost" is captured by parameter "c" in the specification of the analyst's cost of producing information (see eq.(8)). We posit that this cost of multi-tasking increases with the number of stocks followed by an analyst since the number of forecasting tasks for an analyst increases with the number of stocks she covers. Our model predicts that the relation between of data abundance and the informativeness of analysts' forecasts at various horizons should be more pronounced when the cost of multitasking is higher.

[Insert Table VIII about here]

To test this prediction, we reestimate the specifications reported in Table VI interacting each measure of social media data exposure with the number of stocks in analysts' portfolio (i.e., we consider the effect of a triple interaction between the horizon, social media data, and the number of stocks covered by an analyst). Table VIII shows that

the steepening effect of exposure to social media data on the informativeness of analysts' forecasts increases with the number of stocks covered. Consistent with our prediction, all coefficients on the triple interaction term 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 correlation between long-term and short-term earnings. When β is low, our theory predicts that data abundance should have a stronger negative (positive) effect on the informativeness of analysts' long-term forecasts. Intuitively, the reason is that the information collected by an analyst about short-term earnings is less relevant for forecasting long-term earnings when the former are less correlated with short-term earnings.

[Insert Table IX about here]

We test this prediction using firms' earnings auto-correlation as an empirical proxy for β. We obtain it by regressing firms' quarterly earnings on its lag (without a constant) using a rolling window of two years (and requiring at least four observations). Then, we measure β for analyst i on a given day t by the average earnings autocorrelation of all firms covered by the analyst on this day. Finally, we reestimate the specifications reported in Table VI interacting each measure of social exposure data expsure with the average earnings auto-correlation of stocks in analysts' portfolio. Table IX reveals that negative association between analysts' exposure to social media data and the informativeness of their long-term forecasts is less pronounced for analysts covering firms whose earnings are more auto-correlated. The coefficients on the triple interactions are all positive and statistically significant. Thus, in line with our prediction, the steepening of the term-structure of informativeness is weaker when earnings are more auto-correlated.

E.4 Identification Threats

Although our tests provide strong support for the model's predictions, we recognize that there might be alternative explanations for our results. One concern is that our measure of social media data may be correlated with unobserved determinants of analysts' forecasts informativeness that are unrelated to a change in relative costs of collecting

short-term and long-term information (our story). For instance social media activity could be related with news arrival, firm disclosure practices, or their cost of capital, all of which could be associated with the informativeness of analysts' forecasts. Therefore, one may be concerned that such unobserved correlations explain our findings. Alternatively, newly released analysts forecasts may actively foster discussions on social media. In that case, a reverse cauality concerns could arise if new forecasts are differentially informative compared to the previous ones.

We cannot completely rule out such potential alternative stories. Yet, we believe that they are unlikely. Indeed, to jeopardize our interpretation, any candidate explanation should not only explain why forecasts about short-term earnings become more informative as analysts' social media exposure increases, but also why the informativeness of their long-term forecasts simultaneously decreases. In other words, any candidate unobserved variable should (i) correlate *positively* with analysts' exposure to social media data and the informativeness of their short-term forecasts, *and* concurrently (ii) correlate *negatively* with the informativeness of their long-term forecasts.

Furthermore, alternative explanations must also explain our ancillary results. Therefore, any candidate unobserved variable should also be systematically correlated with our proxies for the marginal cost of producing short-term information, the cost of switching tasks, and the auto-correlation of firms' earnings (β). For these reasons, and because it takes a plausible story to invalidate another plausible story, we believe that the observed steepening of the term-structure of analysts' forecasts informativeness associated with their exposure to social media data is likely due to a change in relative costs of collecting short-term and long-term information.

VII Conclusion

This paper examines how data abundance affects the informativeness of financial 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 informativeness 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. If data abundance impairs their long-term forecasts, it might negatively affect the informativeness of asset prices and the efficiency of investment decisions.

References

Abis, Simona, 2018, Man vs machine: Quantitative and discretionary equity management, Discussion paper, .

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

Begeneau, Juliane, Maryam Farboodi, and Laura Veldkamp, 2018, Big data in finance and the growth of large firms, *Journal of Monetary Economics* 97, 71–87.

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).

Crane, Alan, and Kevin Crotty, 2020, How skilled are security analysts?, *Journal of Finance* 75, 1629–1675.

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

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.

Monsell, Stephen, 2003, Task switching, *Trends in Cognitive Science* 7, 134–140.

Myatts, David P., and Chris Wallace, 2012, Endogenous information acquisition in coordination games, *Review of Economic Studies* pp. 340–374.

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.

Womack, Kent, 1996, Do brokerage analysts' recommendations have investment value, *Journal of Finance* 51, 137–167.

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

Figure I: Timeline of the model

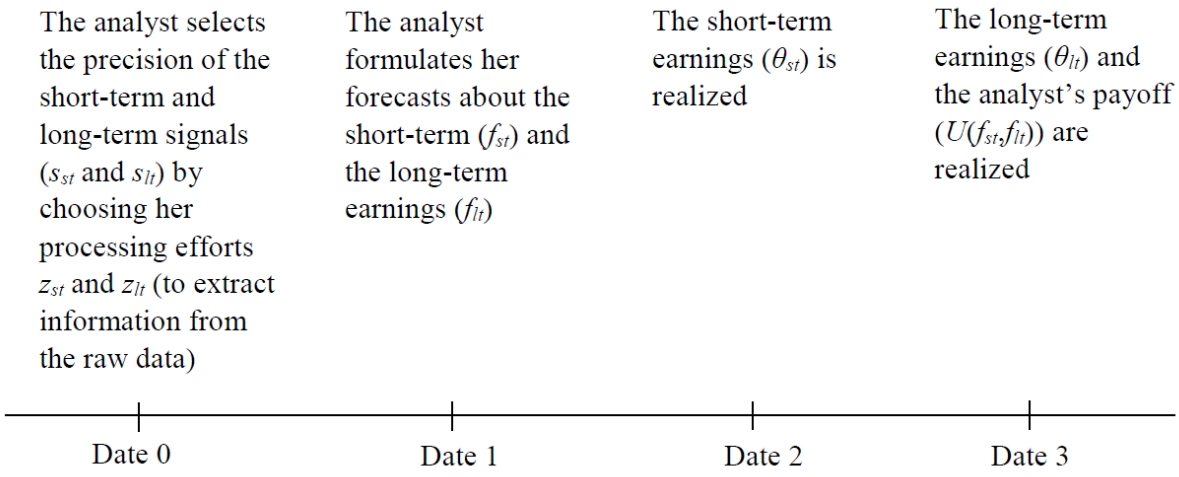
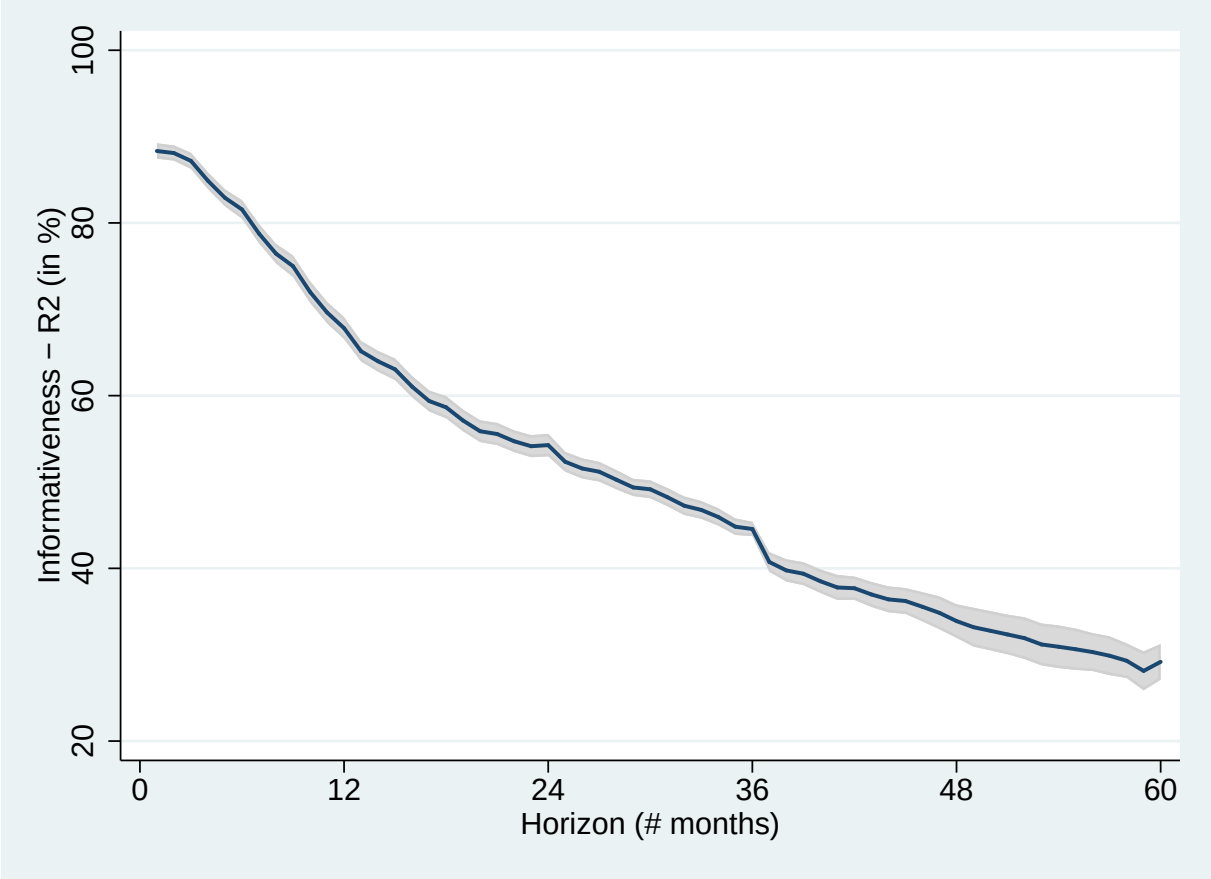
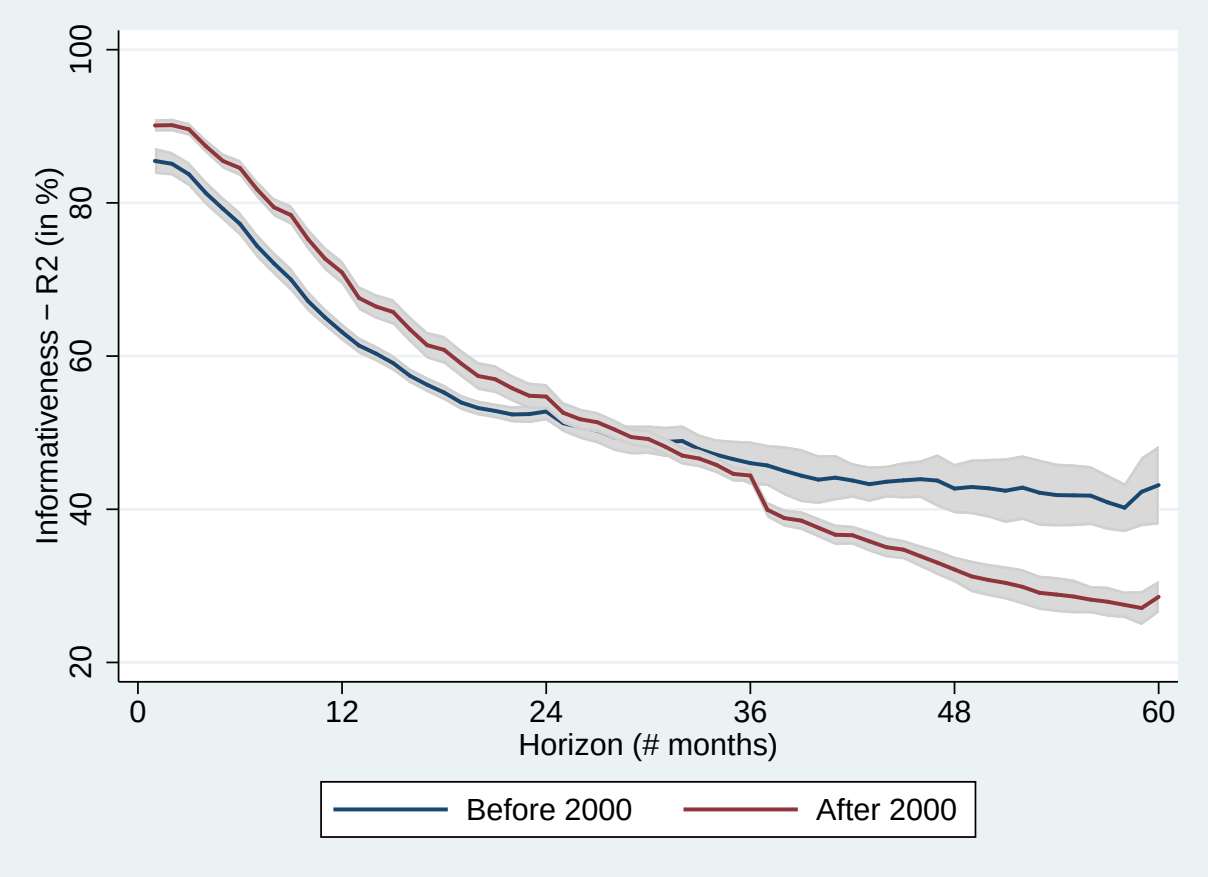


Figure II: The term-structure of analysts forecasts' informativeness



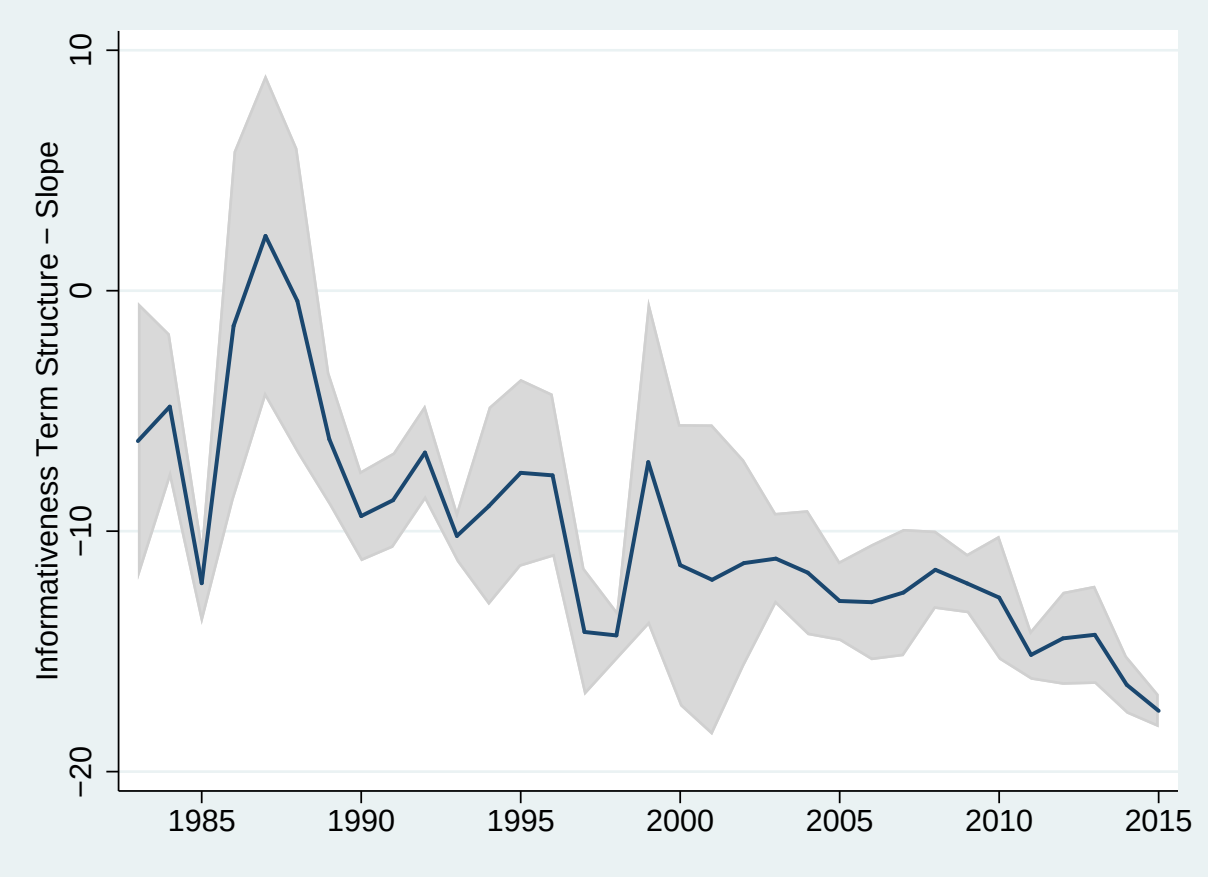
This figure displays the term-structure of analysts forecasts' informativeness. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in months) from zero to five 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



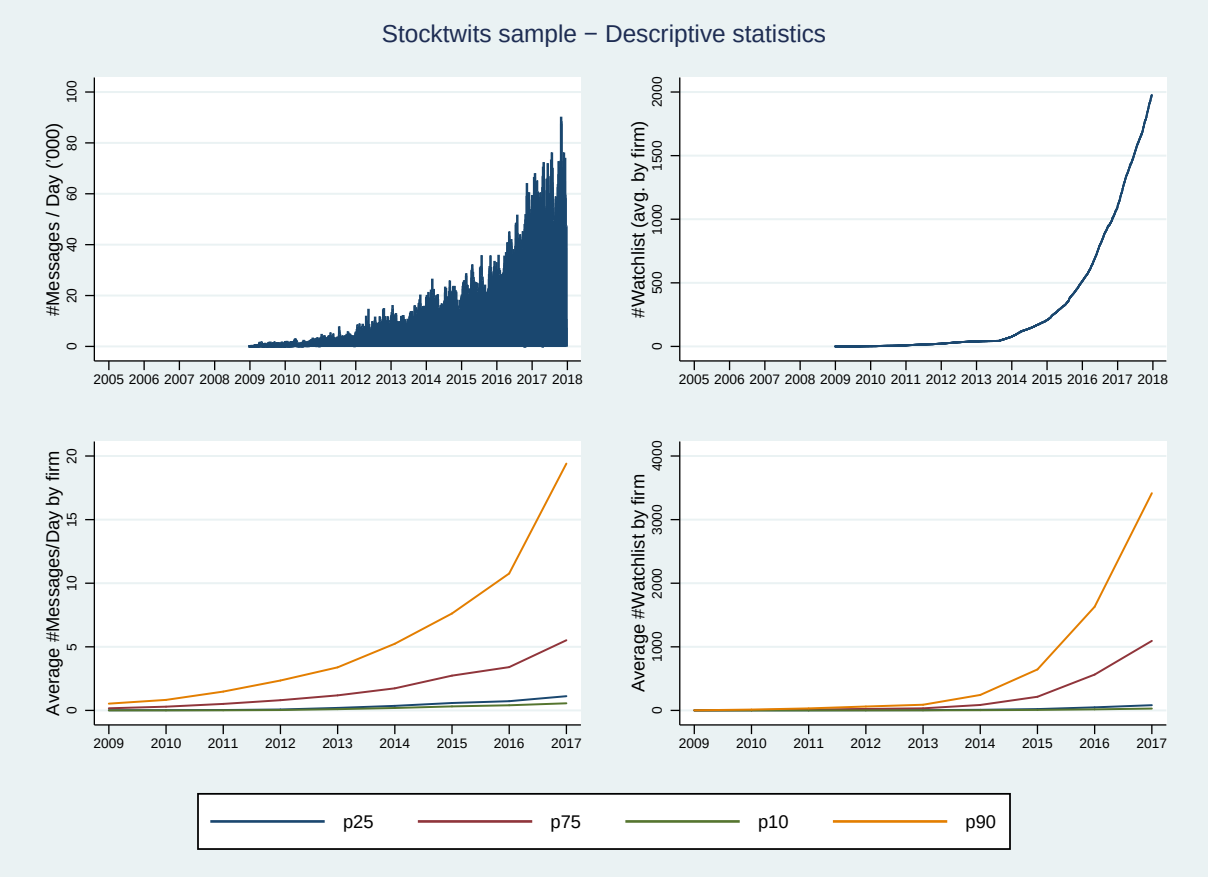
This figure displays the term-structure of analysts forecasts' informativeness before and after 2000. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in 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



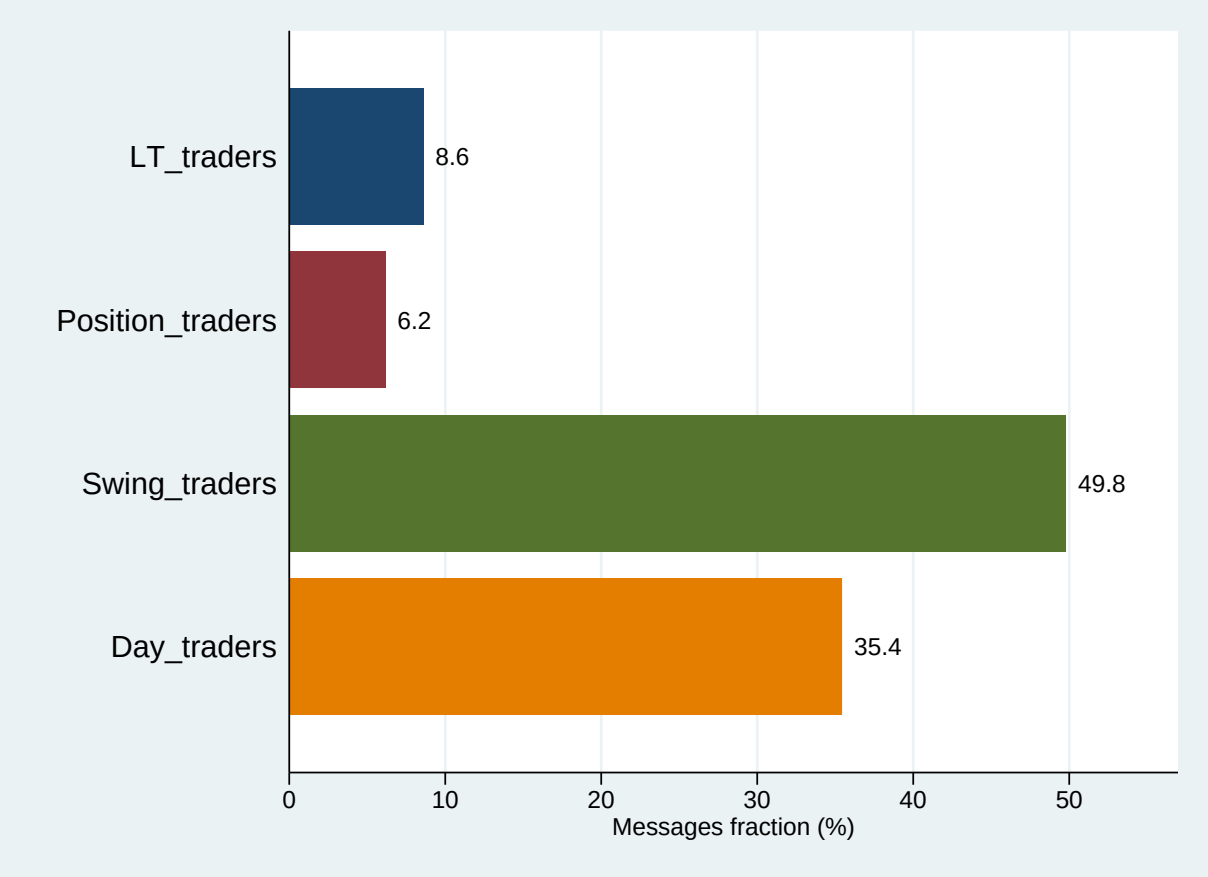
This figure displays the evolution of the slope of the term-structure of analysts forecasts' informativeness. The annual slopes are obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on annual increments of horizon (measured as the 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 used to compute the R^2 measure. 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)	
SIC2 FE	-	No	Yes	-	-
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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. 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 2005-2017 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	0.54***	0.53***	0.4	0.18	-0.65***	-1.00***	-1.51***	-1.55***
	(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	0.47***	0.37***	0.04	-0.32	-0.68	-1.11**	-1.64***	-1.48***
	(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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. 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 2005-2017 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	-0.86***	-0.78***	-0.96***	-0.77***	-0.88***	-0.95***
	(-2.59)	(-3.06)	(-3.72)	(-3.88)	(-4.26)	(-4.31)
Social Media Data	0.13	-0.17	-0.35	0.15	-0.17	-0.39
	(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 2005-2017 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)	-0.50*	-0.41***	-0.46***
	(-1.86)	(-3.98)	(-3.83)
Horizon × (#Messages by Swing Traders)	-0.42	-0.38***	-0.37***
	(-1.46)	(-3.32)	(-3.56)
Horizon × (#Messages by Position Traders)	0.15	0.04	0.07
	(1.14)	(0.35)	(0.70)
Horizon × (#Messages by LT Traders)	-0.02	0	-0.01
	(-0.22)	(-0.04)	(-0.05)
#Messages by Day Traders	0.39	0.23*	0.1
	(1.58)	(1.65)	(0.71)
#Messages by Swing Traders	-0.36	-0.41	-0.46
	(-1.17)	(-1.32)	(-1.59)
#Messages by Position Traders	0.28***	0.14	0.03
	(2.54)	(1.36)	(0.30)
#Messages by LT Traders	0.01	-0.01	-0.01
	(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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. 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 2005-2017 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 Covered	-0.14***	-0.06***	-0.06***	-0.09***	-0.05	-0.05*
	(-5.71)	(-3.38)	(-3.82)	(-5.14)	(-1.52)	(-1.89)
Horizon × Social Media Data	0.69	-0.04	-0.23	-0.05**	-0.03	-0.03
	(1.61)	(-0.10)	(-0.74)	(-2.42)	(-1.10)	(-1.16)
Horizon × #Stocks Covered	-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 Covered	-0.09***	-0.05***	-0.04**	-0.05**	-0.03	-0.03
	(-3.34)	(-2.88)	(-2.26)	(-2.42)	(-1.10)	(-1.16)
#Stocks Covered	-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 Data	1.10***	0.42	0.14	0.76***	0.15	-0.07
	(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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. 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 2005-2017 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 × 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 Data × 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 Data	-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 × Horizon FE		Yes	Yes		Yes	Yes
Date × 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

A Construction of Forecast Informativeness: An Example

This appendix illustrates how the measure of analysts' forecast informativeness, R^2, is computed for a fictitious analyst XYZ covering 6 stocks (A, B, C, D, E, F) on December 31, 2006, and forecasting earnings for the fiscal period ends December 31, 2008. The measurement consists of five steps, illustrated in Table A.1.

- Step 1: Identify the future fiscal period of interest. Since the measure is horizon-specific, forecasts relating to different fiscal periods should not be mixed. In this example we focus on the 2008 fiscal period, and thus ignore the forecasts of XYZ relating to other fiscal periods (e.g., 2007 or 2009).

- Step 2: Retrieve the last available earnings forecast for each covered stock, and the realization of earnings observed ex-post. If the last available forecast is older than 365 days, the analyst is considered inactive on that stock and the R^2 measure is computed excluding that stock.[36] Column 1 of Table A.1 shows the last available earnings forecasts made by XYZ for A, B, C, D, E, and F as of December 31, 2006. The actual realized earnings for fiscal year 2008 are in Column 2.

- Step 3: Normalize earnings. Heterogeneity across firms on size is persistent. To avoid that R^2 reflects that persistence, we normalize both earnings forecasts and realized earnings for each stock by its total assets. Total assets as of December 31, 2008 for A, B, C, D, E and F are in Table A, Column 3. Earnings forecasts (f) and realized earnings (θ) after normalization are reported in Columns 5 and 6.

- Step 4: Estimate R^2 by regressing θ on f in the cross-section of covered stocks (i.e., across A, B, C, D, E and F). R^2 is set to zero if f negatively predicts θ. It is set to missing if there are fewer than 3 or more than 30 observations in the regression, or if the regression coefficient on f is missing after trimming that coefficient at the 1% level in each tail. The R^2 of the regression of θ on f for XYZ on December 31, 2006 is 14.9%.

[36] For example, if as of December 31, 2006, the latest earnings forecast for B made by XYZ were older than 365 days, we would proceed with the R^2 computation without stock B

- Step 5: Compute the horizon defined as the (median) number of days until the earnings realization is publicly released, divided by 365. Column 4 from Table A shows that realized earnings for A, B, C, D, E, and F, were all announced on March 31, 2009. The horizon associated with the above R^2 of 14,9% is thus 2.25 years.

We apply this procedure every day from January 1, 1983 to December 31, 2017 to every US analyst from IBES for all available forecasted fiscal periods. This procedure yields a sample of 65,888,460 daily observations of R^2 with an associated horizon between 1 day and 5 years across 14,379 distinct analysts.

Table A.1: Example of R^2 computation for analyst XYZ on December 31,2006

	Forecasted Fiscal Period: 12/31/2008					
Stock	latest forecast ($million) (1)	realized earnings ($million) (2)	total assets ($million) (3)	earnings report date (4)	latest normalized forecast (f) (5)	realized normalized earnings (θ) (6)
A	110	66	1,100	3/31/2009	0.10	0.06
B	30	18	250	3/31/2009	0.12	0.07
C	59	15	735	3/31/2009	0.08	0.02
D	740	538	6,725	3/31/2009	0.11	0.08
E	1,021	1,225	10,210	3/31/2009	0.10	0.12
F	7	3	55	3/31/2009	0.12	0.06
					R^2	14.9%
					Horizon	2.25

B 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's 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.

C Additional Results

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 Last 30 days		-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 event information at t	No	No	Yes	No
Sample with no event information from t-30 to t	No	No	No	Yes
N	80,434,931	80,379,362	69,414,958	3,147,979

Table A.3: Robustness: Stability of Analysts' Coverage

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 for analysts with stable coverage only. Coverage is stable if the level of similarity between the portfolio of stocks covered in the current year and that of the previous year is greater than 90%. Similarity is defined as the number of common stocks between the portfolio covered in the current year and the one covered the year before, scaled by the square root of the product of the number of stocks in each portfolio. 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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. Social Media Data corresponds to the number of users that have the firm in their watchlist, or to the 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 2005-2017 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 $t-1$. All explanatory variables are normalized by their in-sample standard deviation (except. 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	-0.46	-0.64**	-0.69***	-0.50**	-0.64***	-0.81***
	(-1.49)	(-1.99)	(-2.60)	(-2.43)	(-3.60)	(-3.94)
Social Media Data	0.32	0.08	-0.15	0.33	0.01	-0.18
	(1.25)	(0.33)	(-0.52)	(1.30)	(0.03)	(-0.74)
Horizon	-16.35***			-16.30***		
	(-36.87)			(-35.55)		
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
Low Turnover Sample	Yes	Yes	Yes	Yes	Yes	Yes
N	14,552,054	13,456,617	12,683,350	14,552,054	13,456,617	12,683,350

Table A.4: Robustness: Controlling for Stock Trading Activity

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 for analysts. 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 $t-1$, where t is the date at which we measure forecasts' informativeness. We normalize this variable by its in-sample standard deviation. Social Media Data corresponds to the number of users that have the firm in their watchlist, or to the 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 2005-2017 and both measures of Social Media Data are set to zero prior to Stocktwits introduction in 2009. Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other 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	-1.09***	-1.22***	-1.00***	-1.09***	-1.16***	-1.06***
	(-3.22)	(-3.42)	(-3.74)	(-5.56)	(-6.04)	(-4.96)
Social Media Data	0.16	0	-0.3	0.21	0.03	-0.26
	(0.66)	(-0.02)	(-1.15)	(0.89)	(0.14)	(-1.14)
Horizon × Trading Volume	1.12***	1.19***	0.57***	1.22***	1.31***	0.75***
	(6.56)	(7.58)	(2.67)	(6.92)	(8.14)	(3.84)
Trading Volume	-0.4	-1.19***	-1.23***	-0.43	-1.18***	-1.19***
	(-1.29)	(-3.83)	(-3.80)	(-1.37)	(-4.03)	(-3.93)
Horizon	-17.63***			-17.59***		
	(-31.70)			(-31.18)		
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
Low Turnover Sample	Yes	Yes	Yes	Yes	Yes	Yes
N	30,958,700	28,706,148	27,860,173	30,958,700	28,706,148	27,860,173