Does Alternative Data Improve Financial Forecasting? The Horizon Effect

Olivier Dessaint, Thierry Foucault, and Laurent Frésard*

February 5, 2021

ABSTRACT

We analyze the effect of alternative data on the informativeness of financial forecasts. Our starting hypothesis is that the emergence of alternative data reduces the cost of obtaining information about firms' short-term cash-flows more than their long-term cash-flows. If correct, and forecasting short-term and long-term cash-flows are distinct tasks, analysts will reduce effort to process long-term information when alternative data becomes available. Alternative data thus makes long-term forecasts less informative, while increasing the informativeness of short-term forecasts. We confirm this prediction using variations in analysts' exposure to social media data and a new measure of forecast informativeness at various horizons.

Key words: Alternative data, Security 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 Hazel Hamelin, Gerard Hoberg, Paul Karehnke, Xinyu Liu, Adrien Matray, Randall Morck, Marina Niessner, Gordon Phillips, Jerome Taillard, Laura Veldkamp, and participants at the 2021 AEA Meetings, Copenhagen Business School, the Corporate Finance Webinar, McGill University, NBER Big Data and Securities Markets Conference, INSEAD, the Universita della Svizzera Italiana, and the University of Geneva for useful comments. All errors are the authors' alone. All rights reserved by Olivier Dessaint, Thierry Foucault, and Laurent Frésard.

I Introduction

The digitization of information has generated exponential growth in new types of data available to the financial industry (e.g., data from social media, web traffic, credit card and point-of-sale, geolocation and satellite imagery), often referred to as alternative data.[1] This evolution is transforming the way investors and information intermediaries forecast future outcomes (e.g., cash-flows) and make decisions (e.g., value assets and choose portfolios).[2] However, research on the implications of this transformation is still very limited. In particular, the effects of alternative data on the quality of financial forecasts at different time horizons remain unknown. Our paper addresses this issue, recognizing that many financial decisions (e.g., asset pricing or capital budgeting) rely on forecasts that span both short and long horizons.

More data reduces the cost of information acquisition (Goldfarb & Tucker (2019)). In standard models, a drop in this cost enables forecasters to obtain more precise signals and form more accurate forecasts of asset payoffs (e.g., Verrecchia (1982)). Therefore, we would expect alternative data to improve the quality of financial forecasting, in general. In this paper, we propose a more nuanced prediction. We conjecture that alternative data predominantly contain information about firms' short-term prospects.[3] Alternative data like satellite images of a retailer's parking lots or the number of visits on its website contain information about next quarter earnings but are less clear about earnings in three years.[4] Unlike short-term prospects, long-term prospects depend on firms' strategic and innovation choices. Anticipating these choices and understanding their implications is

[1] According to the website AlternativeData.org, there are more than 400 providers of alternative data in 2020 and the amount invested by buy-side investors in such data is close to $2 billion.

[2] See, for instance, "Demystifying alternative data", Greenwich Associate, 2019 or "How investment analysts became data miners", Financial Times, November 28, 2019.

[3] A brochure from Deustche Bank emphasizes the usefulness of "Estimize" (a social media that crowdsources estimates of future earnings from many individuals) to forecast short-term earnings relative to other sources (See "*The wisdom of crowds: crowdsourcing earnings estimates*", Deustche Bank Market Research, March 4 2014). Specifically, it notes that "*Estimize allows individuals to contribute their estimates anonymously. The underlying concept of the community is to capture the "wisdom of the crowds" in order to reflect investor sentiment and more timely and accurate earnings forecasts.*" Interestingly, it 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 hypothesis.

[4] For evidence that alternative data contains information about short-term firms' earnings, see Froot et al. (2017), Zhu (2019), Katona et al. (2019) and Grennan & Michaely (2019). We are not aware of such evidence for long-term earnings.

difficult without human judgment and expertise (e.g., meetings with industry leaders, scientists, or executives). Therefore, while alternative data can be expected to improve the quality of short-term forecasting, their effect on the quality of long-term forecasts is less clear; it depends on how a decline in the cost of acquiring short-term information affects forecasters' incentive to exert effort to obtain long-term information.

To study this question, we focus on short and long-term earnings forecasts formulated by sell-side equity analysts. We postulate 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. Hence, analysts must strategically allocate effort to produce information at different horizons, like they do when they produce forecasts about different firms (e.g., Harford et al. (2019)).[5] In the presence of multi-tasking costs (due, for instance, to cognitive or time constraints), we show theorically that a drop in the relative cost of obtaining short-term information induces an analyst to allocate more effort to the task of forecasting short-term earnings and less to the task of forecasting long-term earnings hence her short-term forecasts will be more informative, and her long-term forecasts will be less so. Paradoxically, therefore, alternative data could reduce the quality of long-term forecasts though they improve the quality of short-term forecasts. Our study confirms this novel prediction using a new measure of the informativeness of the forecasts of financial analysts and a shock to their exposure to social media data (an important type of alternative data).

To derive this novel prediction, we consider a multitasking problem where one analyst must forecast both the short-term and long-term earnings of a firm. The long-term earnings are the sum of two components: one that is common to short- and long-term earnings (the common component) while the other is unique to long-term earnings (the unique component). The analyst chooses her forecasts to minimize her total expected forecasting error (the weighted average of her expected short- and long-term forecasting errors) net of total information-processing costs. To produce her forecasts, she collects data from which she extracts a short-term signal about the common component and a long-term signal about the unique component. The precision of the signal obtained at

[5]For instance, Harford et al. (2019) postulate that security analysts have "limited time, energy, and resources" and provide evidence that they strategically allocate their limited attention to portfolio firms that matter more to their careers.

a given horizon increases with the effort exerted to process information specific to this horizon. The greater the effort, the higher the marginal cost of improving the precision of the signals at *both* horizons. Thus, the more effort she puts into sharpening the precision of a signal at a given horizon, the greater the cost of further increasing the precision of the other signal. This assumption captures the idea that multi-tasking is costly: giving attention to one task consumes the resources available for others.[6]

Not surprisingly, a reduction in the (marginal) cost of producing the short-term signal induces the analyst to exert more effort to produce this signal. However, albeit optimal, this response raises the marginal cost of producing the long-term signal due to the cost of multi-tasking. As a result, the informativeness of the analyst's long-term forecast drops if the loss in the precision of her long-term signal negates the improvement in the precision of her short-term signal. This happens when (i) the correlation between the short-term and the long-term earnings is sufficiently low, or (ii) the multi-tasking cost is sufficiently high. In both cases, a shock reducing the cost of obtaining short-term information (e.g., the expansion of alternative data), should reduce the informativeness of the analyst's long-term forecasts while improving that of her short-term forecasts.

We test this prediction using the introduction of StockTwits (see below) as a shock that expands the alternative data available to security analysts. To do so we develop a new measure of the informativeness of analysts' forecasts at various horizons that exploits the fact that analysts make recurring earnings' forecasts for multiple firms at different horizons. Specifically, we measure the overall informativeness of the forecasts produced by an analyst on a given day t for a given horizon h (ranging from one day to five years) by the R^2 of a regression of realized earnings at horizon h (across the firms covered by the analyst) on the analyst's forecasts of these earnings. A higher R^2 implies that the forecasts of a given analyst for horizon h explain (in a statistical sense) a larger fraction of the variation in realized earnings for this horizon.[7] In other words, her forecasts reduce uncertainty about future earnings at that horizon for external observers.

[6]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. Hirshleifer et al. (2019) find that an analyst's forecast accuracy declines over the course of a day with the number of forecasts issued by the analyst on this day, due to "decision fatigue".

[7]This measure is similar, in spirit, to the measure of stock price informativeness of Bai et al. (2016). However, it is analyst-specific and we use an analyst's forecasts at a given horizon as explanatory variable for earnings at this horizon (rather than a firm's stock price).

As our measure of forecast informativeness is new, to provide some perspective we begin by studying its evolution in the long run. We use earnings' forecasts made by 14,379 analysts on 13,849 firms between 1983 and 2017 from I/B/E/S (more than 65 million analyst-day-horizon observations). Consistent with existing evidence (e.g., van Binsbergen et al. (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 and four years, and 31.18% for horizons between four and five years. Moreover, between 1983 and 2017, the informativeness of short-term forecasts has increased while that of long-term forecasts has dropped. For instance the informativeness of one-year ahead earnings forecasts has increased by roughly 10 percentage points since 2000 (from about 60% to 70%), whereas the informativeness of five-year ahead forecasts has decreased by roughly 20 percentage points (from more than 40% to less than 30%).

More generally, the annual "slope" of the term-structure of analysts' forecasts informativeness has become significantly more negative over time, both in economic and statistical terms, a trend that has accelerated since 2010, consistent with our main prediction, since the volume of alternative data has grown over time and especially in the last decade. Yet this trend is not direct evidence of our prediction since many other factors such as changes in analysts' compensation packages or changes in corporate disclosure regulations could explain it.

To formally test our prediction, we rely on the introduction (in 2009) of StockTwits, a social networking platform where investors share information (blog posts, charts, or links to articles about a stock) about individual firms.[8] StockTwits has a certain appeal as a laboratory in which to test our hypothesis. Indeed, existing research indicates that financial blog and social media posts (e.g., "Estimize", "MotleyFool", "SeekingAlpha", or "StockTwits") contain information mostly relevant for predicting short-term outcomes

[8]Several recent academic papers use data from StockTwits to measure, for instance, divergence of investors' opinions (e.g., Cookson & Niessner (2020), or Giannini et al. (2019)), the political orientation of their beliefs (e.g., Cookson et al. (2020*a*)), or selective exposure to confirmatory information (e.g., Cookson et al. (2020*b*)). In contrast, we use StockTwits to measure variation in the availability of alternative (social media) data.

(e.g., Chen et al. (2014) or Jame et al. (2016))and the majority of StockTwits users in our sample self-identify as having short-term horizons. Thus, it is likely that StockTwits expands the amount of short-term information available to market participants and thereby reduce the cost of obtaining such information for analysts. In addition, analysts may use data generated by StockTwits since data vendors such as Bloomberg or Thomson Reuters have gradually integrated StockTwits feed on their terminals for market professionals. We show that analysts are more likely to issue a new forecast about a firm when information about this stock is produced on StockTwits.

We estimate how the availability of social media data generated by StockTwits affects the informativeness of analysts' forecasts at different horizons by comparing how the R^2 of an analyst with early and high exposure to discussions on StockTwits changes after 2009 (StockTwits' introduction) relative to that of analysts who are exposed later (or simultaneously) but with less intensity. To isolate the effect of a given analyst's exposure to data generated by StockTwits we consider two daily measures that plausibly (i) capture variations in StockTwits data available to her, and that (ii) would not have been available from other sources in the absence of StockTwits. The first is the daily number of StockTwits' users that have on their "watchlist" the same firms as the ones covered by the analyst. Because users rarely modify their watchlist after registering on StockTwits, the number of users on a firm's watchlist is mostly driven by StockTwits' overall expansion, and is unrelated to the arrival of specific information from other sources (as we confirm empirically). The second is the number of "hypothetical" messages about the firms covered by an analyst posted over the last 30 days. Hypothetical messages on a given firm and day are the total number of messages on StockTwits (for all firms) times the focal firm's historical share of total messages on StockTwits.[9] As the firm's share is persistent, the number of hypothetical messages about the firms covered by an analyst does not change with the arrival of firm-specific information (which we also confirm empirically).

Our test starts in 2005 (i.e., four years before StockTwits) and ends in 2017 so that both measures of analysts' data exposure equal zero before 2009. Using an estimation

[9]Intuitively, "hypothetical messages" is a proxy for the number of messages that would have been observed about a given firm in a given day if, on this day, the messaging activity about that firm relative to other firms was that of a normal day, by historical standards.

approach resembling a standard difference-in-differences (i.e., including analyst and time fixed effects), we exploit heterogeneity in analyst's exposure to StockTwits data (measured as explained above) to analyze how expanding the sources of alternative data affects the term-structure of their forecasts' informativeness.

As predicted, we find that exposure to social media data generated by StockTwits is associated with a significant (i) improvement in the informativeness of analysts' short-term forecasts, and (ii) decline in the informativeness of their long-term forecasts. The economic magnitude of the steepening of the term-structure of analysts' forecast informativeness induced by StockTwits is modest overall. A one standard-deviation increase in an analyst's exposure to StockTwits data increases (decreases) her short (long)-term informativeness by 0.5 (1.5) percentage points. This magnitude is larger for analysts following firms whose users self-identify as having short-term horizons (i.e., day traders and swing traders), that is, when social media data are more likely to contain short-term information. It is also larger for analysts following more firms, for whom the cost of multi-tasking is higher (e.g., Harford et al. (2019)), as well as for analysts following firms whose earnings are less autocorrelated. Thus, the steepening of the term-structure of forecasts informativeness owing to the expansion of alternative data varies systematically, as our theory predicts.

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 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 (e.g., Dugast & Foucault (2018) or Farboodi & Veldkamp (2020)), the growth rates of small and large firms (e.g., Begeneau et al. (2018)), information acquisition choices by asset managers (e.g., Abis (2018)) or the pricing of information by data vendors (Huang et al. (2020)).

Correspondingly, a growing empirical literature has analyzed how reductions in the cost of accessing and producing information due to digitization and the emergence of alternative data affect financial markets and firms' decisions. For instance, Zhu (2019)

and Grennan & Michaely (2019) find that the introduction of satellite images and blog posts has a positive effect on stock price informativeness, while Katona et al. (2019) find no effect of the availability of satellite imagery on price efficiency. Gao & 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) are 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.[10] Goldstein et al. (2020) find that, following the introduction of EDGAR, firms' investment increases, consistent with a decrease in informational asymmetries between firms and investors. However, the sensitivity of corporate investment to stock prices declines, especially for growth firms.

As per this literature, we assume that the emergence of alternative data, combined with progress in computing power, reduces the cost of acquiring information. However, our analysis differs in two important ways. First, we analyze how a reduction in the cost of obtaining short-term information affects the informativeness of forecasts at different horizons when forecasters must allocate their efforts in information acquisition between short-term and long-term cash-flows. To our knowledge, this issue has not been explored in the literature.[11] Yet, it is relevant because most financial decisions require making forecasts about outcomes that 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 central information intermediaries whose output is informative and affects stock prices (see, for instance, Womack (1996), Merkley et al. (2017) or Crane & Crotty (2020)). For this reason, understanding how alternative data affect the informativeness of their forecasts is important if we are to discern the channels through which alternative data might affect stock price informativeness (e.g., a drop in the informativeness of analysts' long-term forecast can reduce the informativeness of stock prices for long-term cash-flows). Our findings thus add to the literature on the determinants of analysts' effort allocation (e.g., Harford et al. (2019) or Hirshleifer et al. (2019)),

[10]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.

[11]Dugast & Foucault (2018) shows that a decrease in the cost of producing signals after new information arrival strengthens the informativeness of stock prices in the short-term but not necessarily in the long-term, where short-term and long-term are defined by the time elapsed *since* news arrivals. This is distinct from the notions of short-term and long-term used in our paper (the time elapsed *until* the realization of a payoff).

the properties of short-term and long-term forecasts (e.g., Bandyopadhyay et al. (1995) or Mest & Plummer (1999)), and how progress in information technologies affects the organization and output of security analysts (e.g., Grennan & Michaely (2020) or van Binsbergen et al. (2020)).

III Hypothesis Development

Our core new idea is that forecasting earnings at different horizons are distinct tasks to which analysts strategically allocate effort in order to produce information. In the face of multi-tasking costs, a shock to the cost of obtaining information at a specific horizon (e.g., when data relevant for this horizon emerges) will affect the allocation of effort between all horizons, and in turn the entire term-structure of analysts' forecast informativeness.

A The Analyst's Forecasting Problem

Figure I presents the timeline of the model. One firm generates 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, the analyst announces her forecasts for these earnings. The short-term earnings are normally distributed with mean zero and variance σ^2_{st}. The long-term earnings are:

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

where e_{lt} is normally distributed with mean zero and variance σ^2_e and is independent from θ_{st}. Thus, the long-term earnings are the sum of: (i) one component that depends on the short-term earnings, and (ii) one component orthogonal to the short-term earnings. One interpretation is that the first component is generated by assets in place and the second by growth opportunities (e.g., due to innovations). The correlation between the long-term and the short-term earnings increases with β.

[Insert Figure I about here]

Let f_{st} and f_{lt} be, respectively, the short-term forecast (about θ_{st}) and the long-term forecast (about θ_{lt}) of the analyst. The analyst's payoff $W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st})$, which is realized at date 3, is inversely related to the weighted sum of her 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 $\gamma \in [0, 1]$. This specification captures the fact that analysts' career outcomes (compensation and upward mobility) are positively related to the accuracy of their forecasts. For instance, Hong & Kacperczyk (2010) and Harford et al. (2019) show that analysts with more accurate forecasts are more likely to be ranked "all star analysts" or be promoted. When γ increases, the analyst's payoff becomes more sensitive to the quality of her short-term forecasts and less sensitive to the quality of her long term forecasts.

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

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

where $\Omega_1 = \{s_{st}, s_{lt}\}$ are the signals used by the analyst to formulate her forecasts at date 1. These are extracted from raw data (e.g., accounting data, analysts meetings, industry reports, regulatory filings, news and scientific articles, social media) collected and processed by the analyst at date 0. The first (the "short-term signal"), s_{st}, is about the common component (θ_{st}) of the short-term and long-term earnings, while the second (the "long-term signal"), s_{lt}, is about the unique component (e_{lt}) of the long-term earnings:

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

where the εs are the noise in the analyst's signals. These variables are normally distributed and independent from all other random variables in the model.

Given her signals, the analyst optimally chooses her forecasts at date 1 to solve:

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

Using eq.(3), we obtain that the analyst's optimal forecasts for the short-term and long-term earnings are as follows (all derivations are provided in Appendix A):

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

We deduce from eq.(3) and eq.(6) that the analyst's expected payoff *at date* 0 is:

$$\begin{aligned}\mathsf{E}(\bar{W}(f^*_{st}, f^*_{lt}; s_{st}, s_{lt})) &= \omega - \gamma\,\mathsf{E}((\mathsf{E}(\theta_{st}\,|s_{st}) - \theta_{st})^2) - (1-\gamma)\,\mathsf{E}((\mathsf{E}(\theta_{lt}\,|s_{st}, s_{lt}) - \theta_{lt})^2),\\ &= \omega - \gamma\,\mathsf{E}(\mathsf{Var}(\theta_{st}\,|s_{st})) - (1-\gamma)\,\mathsf{E}(\mathsf{Var}(\theta_{lt}\,|s_{lt}, s_{st})),\\ &= \omega - \underbrace{q(\beta,\gamma)\,\mathsf{Var}(\theta_{st}\,|s_{st}) - (1-\gamma)\,\mathsf{Var}(e_{lt}\,|s_{lt})}_{\text{Total Expected Forecasting Error}},\end{aligned} \tag{7}$$

where $q(\beta,\gamma) \equiv (\gamma + (1-\gamma)\beta^2)$.[12] Thus, the analyst's *expected* payoff increases when her signals are less noisy because her total expected forecasting error is then smaller.

To reduce the noise in her signals, the analyst can exert an effort to obtain more information about the common or the unique components of the firm's future earnings. We see this problem as two different tasks requiring specific efforts, very much like formulating forecasts about different – but related – firms in an analyst's portfolio are different tasks (e.g. Harford et al. (2019)). To capture this idea, we assume that $\epsilon_h \sim \mathcal{N}(0, (Z - z_h)\xi_h^2)$ where $z_h \leq Z$ is the analyst's effort to collect and process information at horizon $h \in \{st, lt\}$.[13] Thus, the precision of the analyst's signal s_h is $\tau_h(z_h) = ((Z - z_h)\xi_h^2)^{-1}$, for $h \in \{st, lt\}$ and the larger is the analyst's effort, z_{ht}, to produce the signal at horizon h, the higher is the precision of this signal. If the analyst does not make an effort for producing a signal at horizon h, she obtains a signal of precision $(Z\xi_h^2)^{-1}$ (which goes to zero when Z becomes large).

The analyst's total cost of effort to generate the short-term and the long-term signals is:

$$C(z_{st}, z_{lt}) = C_0 + az_{st}^2 + bz_{lt}^2 + cz_{st}z_{lt}, \tag{8}$$

where $C_0 \geq 0$ is the fixed cost of understanding the firm's business and collecting information about it (a prerequisite for coverage). We assume that $a > 0$ and $b > 0$: the marginal cost of effort to improve the precision of a signal at a given horizon increases with the level of effort. We further assume that $c > 0$, that is, multi-tasking is costly for the analyst because forecasting the common and the unique components of short-term

[12] The last line in eq.(7) follows from the fact that (i) $\mathsf{Var}(\theta_{ht}\,|s_{ht})$ does not depend on the realization of s_{ht} because θ_{ht} and s_{ht} are normally distributed, and (ii) the independence between the common component (θ_{st}) and the unique component (e_{lt}) in the long-term earnings.

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

and long-term earnings are different tasks. The former requires understanding the firm's assets in place, whereas the latter requires understanding the firm's growth opportunities. If the analyst already exerted a lot of effort to improve the precision of, say, her short-term signal then it becomes more demanding for her to exert even more effort, be it to improve the precision of the short-term signal even further ($a > 0$) or the precision of the long-term signal ($c > 0$). One can also interpret c as capturing switching costs associated with multi-tasking: If the analyst devotes much time to forecasting the common component of short-term and long-term earnings then switching to the task of forecasting the unique component is costly and vice versa.[14]

The analyst chooses her efforts, z_{st}^* and z_{lt}^*, at date 0 to maximize her ex ante expected payoff *net* of information processing costs. That is, z_{st}^* and z_{lt}^* solve

$$\begin{aligned} Max_{\{z_{st},z_{lt}\}\in[0,Z]^2} J(z_{st},z_{lt}) &= \mathsf{E}(\bar{W}(f_{st}^*,f_{lt}^*;s_{st},s_{lt})) - C(z_{st},z_{lt}) \\ &= \omega - \underbrace{q(\beta,\gamma)\,\mathsf{Var}(\theta_{st}\,|s_{st}) - (1-\gamma)\,\mathsf{Var}(e_{lt}\,|s_{lt})}_{\text{Total Expected Forecasting Error}} - C(z_{st},z_{lt}), \end{aligned} \tag{9}$$

Thus, in choosing her efforts, the analyst trades-off accuracy (more effort at a specific horizon reduces the total expected forecasting error) against the cost of effort.[15]

To solve for the analyst's optimal efforts, it is analytically convenient to assume that the analyst's priors about θ_{st} and e_{st} are diffuse. This simplifies expressions for the analyst's expected forecasting errors and the derivations. Indeed, in this case, as all variables are normally distributed, we obtain:[16]

$$\begin{aligned} \mathsf{Var}(\theta_{st}\,|s_{st}) &= (Z - z_{st})\xi_{st}^2, \\ \mathsf{Var}(e_{lt}\,|s_{lt}) &= (Z - z_{lt})\xi_{lt}^2. \end{aligned} \tag{10}$$

Substituting these expressions in eq.(9), we obtain the following result.

[14]Switching costs associated with multi-tasking are well documented in the psychological literature. See, for instance, Monsell (2003).

[15]We assume that the analyst is always better off covering the firm, i.e., $J(z_{st}^*, z_{lt}^*) \geq 0$. As the analyst can always choose to exert no effort ($z_{st} = z_{lt} = 0$) and obtain the signals s_{st} and s_{lt} with minimal precision, a sufficient condition for this is that $J(0,0) \geq 0$, which is satisfied for ω large enough.

[16]As θ_h and s_h are normally distributed, $\mathsf{Var}(\theta_h\,|s_h)$ is equal to the inverse of the sum of the precision of θ_{st} and the precision of s_{ht}. When priors are diffuse, the precision of θ_{st} is zero and $\mathsf{Var}(\theta_h\,|s_h)$ is simply the inverse of the precision of s_{ht}. Eq.(10) follows.

Proposition 1 *: When $c \leq \bar{c}(\beta, \gamma, a, b, \xi_{st}^2, \xi_{lt}^2)$ (where $\bar{c}$ is defined in the proof of the proposition) and Z is large enough, the analyst's optimal efforts in producing information at date 0, z_{st}^* and z_{lt}^*, are interior (i.e., $\{z_{st}^*, z_{lt}^*\} \in [0, Z]^2\}$) and given by:*

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

Thus, z_{st}^* decreases with the marginal cost of producing the short-term signal ("a") while, if $c > 0$, z_{lt}^* increases with this cost.

As explained in the introduction, our hypothesis is that alternative data predominantly contain short-term information and therefore reduce the cost of obtaining short-term information (a in our model), i.e., about the common component of short and long-term earnings. In turn, the drop in cost leads the analyst to put more effort in improving the precision of the short-term signal ($\frac{\partial z_{st}^*}{\partial a} < 0$). Less obviously, when $c > 0$, it also leads the analyst to *reduce* her effort to improve the precision of her long-term signal ($\frac{\partial z_{lt}^*}{\partial a} > 0$ iff $c > 0$) because, as the analyst optimally puts more effort into processing and collecting short-term information, the marginal cost of doing so for long-term information rises. In other words, when the cost of obtaining short-term information drops, the analyst optimally focuses on improving the precision of her short-term signal while cutting efforts to produce the long-term signal to save on multi-tasking costs. This mechanism is key to our main predictions, as explained below.

Remark: In the Internet Appendix, we show that if $c \leq \frac{4C_0}{q(\beta,\gamma)(1-\gamma)\xi_{st}^2\xi_{lt}^2}$ then it is not optimal for the analyst (or the intermediaries employing the analyst) to divide the tasks of producing the short and long-term signals between two agents (to save on the cost of multi-tasking, c). Indeed, under this condition, the increase in fixed costs of information production (each agent bears the fixed cost C_0 of collecting information to understand the firm's business) cancels out the drop in the cost of multi-tasking. Other frictions (communication and agency; see the internet appendix for an analysis) can also explain why splitting the tasks of forecasting short-term and long-term cash-flows between two agents might not be optimal in reality, even when $C_0 = 0$ (see Goldstein & Yang (2015) for a similar point in a different context).[17]

[17] Goldstein & Yang (2015) consider model in which the payoff of an asset is the sum of three components

B Alternative Data and Analyst 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 $h \in \{st, lt\}$, denoted by $\mathcal{I}_h$, as the inverse of the variance of the firm's realized earnings at this horizon conditional on the analyst's forecast at this horizon.[18] That is:

$$\mathcal{I}_h \equiv \mathsf{Var}(\theta_h \mid f_h^*)^{-1} \quad for \quad h \in \{st, lt\} \tag{12}$$

As $f_{st}^* = \mathsf{E}(\theta_{st} \mid s_{st})$ and $f_{lt}^* = \mathsf{E}(\theta_{st} \mid s_{st}, s_{lt})$, we have (see Appendix A):

$$\mathcal{I}_{st} = \mathsf{Var}(\theta_j \mid s_{st})^{-1} = ((Z - z_{st}^*)\xi_{st}^2)^{-1}, \tag{13}$$

and

$$\mathcal{I}_{lt} = \mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})^{-1} = (\beta^2((Z - z_{st}^*)\xi_{st}^2) + (Z - z_{lt}^*)\xi_{lt}^2)^{-1}. \tag{14}$$

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

Corollary 1 *: If $\beta < (\frac{c}{2b})^{\frac{1}{2}} \frac{\xi_{lt}}{\xi_{st}}$, a decrease in the marginal cost of producing the short-term signal (a) triggers an increase in the informativeness of the analyst's short-term forecast and a decrease in the informativeness of the analyst's long-term forecast.*

As explained previously, a decrease in the marginal cost of producing the short-term signal (a) results in a reallocation of the analyst's effort: she puts more effort into increasing the

and investors can acquire information about the first component or the second one or both. They assume that the cost of acquiring information on both components is higher than the sum of the costs of acquiring information on each component separately. This assumption is similar to our assumption that $c > 0$. In their setting as well, delegating information acquisition on both components to two agents would be less costly than information acquisition by a single agent. However, Goldstein & Yang (2015) write (p.1747):"*An intermediary can hire two such traders [...]. In our view, however, it is unlikely that such combinations happen without any friction, as agency costs, coordination costs or organizational costs imply that combining two traders with different pieces of information will be an additional cost*".

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

precision of the short-term signal ($\frac{\partial z^*_{st}}{\partial a} < 0$) and less effort into increasing the precision of the long-term signal ($\frac{\partial z^*_{lt}}{\partial a} > 0$). The first effect raises the informativeness of the long-term forecast (because it enables the analyst to better forecast the common component of the firm's future earnings) while the second effect reduces it (because it impairs the analyst's ability to forecast the unique component of the long-term earnings). As shown by Corollary 1, the second effect dominates when long-term earnings are not closely correlated with short-term earnings or when the cost of multi-tasking is large enough ($\beta < (\frac{c}{2b})^{\frac{1}{2}} \frac{\xi_{lt}}{\xi_{st}}$).[19] In this case, the informativeness of the long-term forecast declines with the cost of producing short-term information. In contrast, the informativeness of the short-term forecast always improves.[20]

These differential effects of a drop in the marginal cost of producing the short-term signal lead to our main testable hypothesis. Insofar as the emergence of alternative data reduces relatively more the cost of obtaining short-term information, this evolution should be associated with (i) an increase in the informativeness of analysts' short-term forecast (as one would expect), and (ii) a decrease in the informativeness of their long-term forecasts, especially for firms with a relatively low auto-correlation of earnings (low β) or analysts for which the cost of multi-tasking (c) is high. Although both predictions are interesting, the second is less obvious than the first, and more specific to the multi-tasking channel highlighted by our model (the assumption that $c > 0$). In the rest of the paper, we test these predictions and find support for them.

IV Data and Measurements

A Earnings Forecasts and Realizations

To measure analysts' forecasts informativeness at different horizons, we first build a large sample of earnings forecasts 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

[19] Note that the condition $\beta < (\frac{c}{2b})^{\frac{1}{2}} \frac{\xi_{lt}}{\xi_{st}}$ requires either β low enough or c high enough. Existence of an interior solution to the analyst's problem requires $c < \bar{c}$ (see Proposition 1). Using the expression for $\bar{c}$ given in the appendix, it can be checked that the set of parameter values (e.g., for c and β) such that these two conditions hold is non empty.

[20] The same predictions hold for the *joint* informativeness of analysts' forecasts, i.e., $\mathcal{I}^{joint}_{ht} = \mathsf{Var}(\theta_{lt} \,|f^*_{st}, f^*_{lt})^{-1}$ for $h \in \{st, lt\}$ because the analyst's optimal forecast of the earnings at horizon h is equal to her expectation of these earnings conditional on her signals. This implies, for instance, $\mathcal{I}^{joint}_{lt} = \mathsf{Var}(\theta_{lt} \,|f^*_{st}, f^*_{lt})^{-1} = \mathsf{Var}(\theta_{lt} \,|s_{st}, s_{lt})^{-1} = \mathcal{I}_{lt}$.

and Unadjusted). We exclude quarterly and semi-annual earnings forecasts, and retain annual earnings forecasts associated with a clearly defined fiscal period.[21] 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 use net income forecasts as our main measure of "earnings" forecast.[22] 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 or otherwise the number of shares from CRSP. Then, to obtain our final sample of earnings forecasts, we apply the following criteria. All earnings forecasts must be about a fiscal year ending between 1983 to 2017. We require that actual earnings for the forecasted fiscal period and total assets from Compustat at the end of the forecasted fiscal period be available. 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 and we impose that actual net income (in absolute value) is not greater than total assets at the end of that period.

[21] 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.

[22] 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. This conversion is not trivial because I/B/E/S does not report the number of shares used by the analyst to make the EPS forecast. We find that the best approach to minimize the risk of error is to multiply the actual net income 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.

B Measuring Forecasts Informativeness

Using this sample of earnings forecasts and realizations, we then construct a daily measure of forecast informativeness by analyst and forecasting horizon. The horizon 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. Appendix B illustrates how we compute this measure on a given day for one analyst and one specific horizon. To compute that measure daily for all analysts and every possible horizon, we proceed as follows.

First, we retrieve all earnings forecasts most recently issued by an analyst for a specific (future) fiscal period (hereafter the forecasted fiscal period). Next, for each analyst and forecasted fiscal period, we create a firm-day panel including all forecasts issued by the analyst for that fiscal period. Hence, this panel is specific to an analyst and forecasted fiscal period. It starts on the date of the first forecast made for that fiscal period and ends when the firms covered by the analyst announce their earnings.[23] Since an analyst does not update her forecasts daily, this panel has gaps, which we fill using her last available forecast.[24] Doing so allows us to identify forecasts most recently issued by this analyst on every date of the panel, to which we associate a unique measure of horizon. Intuitively, horizon decreases incrementally each day by one day. Therefore, we define horizon on a given date of the panel as the number of days until actual earnings are announced. Since earnings announcement dates generally differ across firms covered by an analyst, we compute the median date and define the horizon as the number of days until that median date, divided by 365. At the end of this process, a given analyst-day-horizon assembles a collection of forecasts issued by a given analyst about the firms she covers for a given forecasting horizon.

We define the informativeness of the forecasts of analyst (i) on day (t) for horizon (h) as the R^2 of the following regression:

$$e_j = k_0 + k_1 \hat{e}_j + \nu_j, \tag{15}$$

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

[24] To avoid stale forecasts, we only consider the last available forecast if it is not older than one year.

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, respectively, the forecasted and realized earnings for firm j normalized by total assets at the end of the forecast period.[25] By definition, the R^2 of this regression is:

$$R^2_{i,t,h} = 1 - \frac{\mathsf{Var}(\nu_j)}{\mathsf{Var}(e_j)} = 1 - \frac{\mathsf{Var}(e_j|\widehat{e_j})}{\mathsf{Var}(e_j)}. \tag{16}$$

A higher $R^2_{i,t,h}$ implies that analyst i's forecasts of earnings at horizon h on day t explains a larger fraction of the variation in realized earnings at date $t+h$ across the firms she covers. In this sense, a higher $R^2_{i,t,h}$ indicates that analyst i's forecasts at horizon h are more informative. As explained below, this measure is closely related to the theoretical measure of forecasts informativeness analyzed in Section III.B.

We obtain $R^2_{i,t,h}$ by estimating eq.(15) for each available analyst-day-horizon collection for horizons varying from one day to five years. For estimating eq.(15), we require (a) more than three, but (b) less than thirty forecasts observations at horizon h by a given analyst on day t so that (a) $R^2_{i,t,h}$ can be estimated, and (b) we avoid using forecasts issued by teams rather than individual analysts. Finally, to limit the effect of outliers coming from lower power in estimations with few observations, we drop estimates of $R^2_{i,t,h}$ when the estimated slope of eq.(15) (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.

C Why Using R^2 as a Measure of Forecast Informativeness?

Our measure of the informativeness of an analyst's forecast (R^2) is similar in spirit to the measure of stock price informativeness developed by Bai et al. (2016).[26] However, it is new in the literature on analysts, which generally uses two measures of analysts' forecast informativeness: (i) the absolute (or squared) forecast error (often called accuracy) or (ii) the impact of analysts' forecasts on stock prices (e.g., Hilary & Hsu (2013) or Merkley et al. (2017)).

[25] Our results are identical when normalizing by total assets from the last available financial statements on day t. One drawback with this alternative approach is that the measure of informativeness can change even when analysts do not update their forecasts (because the normalization changes).

[26] Bai et al. (2016)'s measure of price informativeness is based on yearly cross-sectional regressions of firms' future earnings at various horizons on firms' market-to-book ratios.

Several reasons motivate our choice. First, our measure of forecasts' informativeness is very close to our theoretical measure, $\mathcal{I}$ (see eq.(12)). Indeed, the theoretical R^2 of a regression of θ_h on f_h in the model is $R^2_h = 1 - \mathsf{Var}(\theta_h \mid f_h)/\mathsf{Var}(\theta_h)$ for $h \in \{st, lt\}$. Thus, when the informativeness of the analyst's forecast is higher in theory (i.e, when $\mathsf{Var}(\theta_h \mid f_h)$ is smaller), R^2_h is higher. One advantage of using R^2_h is that this measure accounts for the intrinsic difficulty of the forecasting exercise by normalizing $\mathsf{Var}(\theta_h \mid f_h)$ by $\mathsf{Var}(\theta_h)$. Indeed, our theoretical measure of forecast informativeness can vary either because the analyst's efforts in producing information vary or because uncertainty about earnings in the absence of effort (captured by $Z\xi_h$ for earnings at horizon $h \in \{st, lt\}$ in the model) varies. We are interested in measuring the first source of variation, not the second one. It is worth stressing that $R^2_{i,t,h}$ is specific to an analyst in the data, not specific to an analyst and a firm as in the model. In effect, we are treating each pair $(e_j, \hat{e}_j)$ at a given horizon h for firm j covered by analyst i as different realizations of the pair (θ_h, f^*_h) for $h \in \{st, lt\}$ in the model. By estimating eq.(15) by ordinary least squares, our implicit assumption is that firms' normalized earnings (e_j) and forecasts $(\hat{e}_j)$ at various horizons in a given analyst's portfolio are realizations of the same underlying (gaussian) distributions.[27]

Second, absolute (or squared) forecast errors at a given horizon can be large because an analyst exerts little effort to collect information relevant for this horizon or because uncertainty at this horizon is large (or both). As just explained, we are interested in measuring the former effect, not the latter. This is better achieved by normalizing the measure of analyst's informativeness by a measure of uncertainty about future earnings. Our R^2 measure naturally achieves this because, in R^2, the explained variance of future earnings at a given horizon, $\mathsf{Var}(e_j|\widehat{e_j})$, is divided by the (cross-sectional) unconditional variance of firms' realized earnings $\mathsf{Var}(e_j)$.

Third, the literature has found that analysts' forecasts are positively biased (they are too optimistic), maybe due to conflicts of interest (e.g., Hong & Kacperczyk (2010)). Analysts' squared (or absolute) forecasts errors increase in the average level of the bias and its dispersion across firms covered by the analyst. In contrast, our measure of analyst's forecast informativeness is not affected by the average level of the analyst's bias and

[27] As analysts tend to follow firms with similar product market characteristics, heterogeneity across firms in an analyst's portfolio is usually low (especially after normalizing by firms' size as we do).

identical to the informativeness of analysts' debiased forecasts if the analyst's bias is constant across firms covered by the analyst at a given point in time (see the Internet Appendix for a formal proof). Thus, our measure of analysts' forecasts informativeness is less likely to be affected by determinants of analysts' biases than analysts' absolute or squared forecasts errors

Last, measuring the informativeness of analysts' forecasts by their impact on stock prices is problematic for our purpose. Indeed, analysts often issue long-term forecasts only in association with short-term forecasts. This simultaneity makes it impossible to build a market-based measure of forecast informativeness for a specific *horizon* because one cannot measure the extent to which each forecast at a specific horizon contributes to observed price reactions (forecasts at all horizons matter for stock prices; e.g., Bandyopadhyay et al. (1995) or Mest & Plummer (1999) make a similar point).

D The Term-Structure of Forecasts' Informativeness

Table I presents summary statistics for our measure of analysts' forecasts informativeness. Across all horizons, the average informativeness of analysts' forecasts is 68.01%. The average analyst's earnings forecasts explain 68% of the variation in realized earnings across the firms she covers. We note a substantial variation in R^2 across analysts, with a sample standard-deviation of 33.90%. On average, an analyst covers 8.12 firms on any given day. 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 factors contribute to this asymmetry. First, although all analysts have to forecast short-term and long-term earnings to value firms and make investment recommendations, they 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 earnings are realized (e.g., through acquisitions).[29]

[28]The prevalence of short-term forecasts thus likely reflects I/B/E/S selected information retrieval from analysts' reports and/or analysts' reporting choice. This later case could raise potential concerns for our analysis if analysts' disclosure depends on their intrinsic forecasting quality (e.g., only analysts of lower quality report their long-term forecasts). We significantly lessen such concern in later tests by focusing specifically on a subset of analysts reporting both short-term and long-term forecasts (e.g., Panel B Table II), and controlling for analyst fixed effects (e.g., column 5 from Table III, and Table V and after).

[29]We also do not observe earnings' realizations associated with long-term forecasts issued towards the end of our sample, because these earnings have not realized yet.

[Insert Table I and Figure II about here]

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 for horizon less than one year compared to 6.70 for horizons ranging between three and four years.[30]

Table I also reveals that the informativeness of analysts' forecasts varies significantly by horizon. The average R^2 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 on dummy variables capturing each (daily) horizon (from one day to five years). Figure II plots the estimated coefficients (together with 90% confidence intervals) and confirms that forecasts at shorter horizons are significantly more informative than forecasts at longer horizons.

Figure II (and Table I) suggests that the informativeness of forecasts decays quickly with their horizon. Indeed, analysts' forecasts are about two times more informative at the one-year horizon than at the five-year horizon. A linear approximation obtained by regressing R^2 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 each one-year increase in the forecasting horizon.

V Long-Term Evolution

Clearly, the volume and variety of available alternative data has increased over time. For instance, Figure III displays the number of alternative data providers reported by alternativedata.org, a practitioners' website tracking available alternative data sources (mostly for institutional investors).[31] The number of providers increased from less than 50

[30] Thus, our estimates of R^2 at longer horizons are less precise since they are obtained from estimations of eq.(15)) with fewer observations.

[31] Available at: https://alternativedata.org/stats/.

in the late nineties to more than 400 in recent years. Our hypothesis is that this evolution has predominantly reduced the cost of accessing short-term information. According to our model, this should lead analysts to focus more attention to the production of short-term information, possibly at the expense of the production of long-term information. Thus, our hypothesis predicts a steepening of the term-structure of analysts' forecasts informativeness over time. Our goal in this section is to check whether this trend is present in the data.

This check is however not a test of our predictions, since it could be argued that many other factors might have affected the trend in the term-structure of analysts' forecasts informativeness (see below for more discussion). However, this trend provides the background for our tests in the next section and is informative about long-term trend in firms' informational environments. In particular, it complements recent analyses of long-term trends in stock price informativeness (e.g., Bai et al. (2016), Farboodi et al. (2020) or Davila & Parlatore (2020)) motivated by changes in information and trading technologies by considering another source of information for investors (namely, analysts' forecasts).

[Insert Figure III about here]

Figure IV shows 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 half of our sample.[32] To formally test whether this shift corresponds to a general trend, 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 forecasts' informativeness over the 1993-2017 period.[33]

[Insert Figure IV and Table II about here]

[32]Relatedly, using the precision of the idiosyncratic information component in forecasts to measure their quality, Srinidhi et al. (2009) report that the quality of short-term forecasts improved in the two years following the enactment of regulation Fair Disclosure (FD) in 2000 (compared to two years prior), whereas that of long-term forecasts declined.

[33]In this test, we cluster standard errors by forecasted fiscal period. Changing the level of clustering does not materially affect statistical inference.

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). However, Columns (7) and (9) indicate that the informativeness of long-term forecasts (more than three and four years) has deteriorated, with coefficients on the trend of -11.5 and -20.

Panel A further reports specifications that include fixed effects for two-digit SIC industries (corresponding to the main industry covered by each analyst and year), as well as fixed effects for the average (quintile) 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 restrict our estimation to analysts issuing forecasts at both short and long horizons and find a similar shift in the term-structure.

[Insert Figure V 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 year by year starting in 1983. Figure V reveals that it has become significantly steeper (i.e., more negative) over time, and even more so in recent years. While it remained above -10 until the mid-90s, its steepening accelerated over the second half of our sample period, especially after 2005. After this date, the slope is consistently and (statistically) significantly lower than -10. Table III confirms this pattern. We regress the slope estimates on a normalized trend with annual increments starting in 1993, as we did in Table II. 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 (coefficient of -10.6 with a t-statistics of -6.26) to reach a slope of -17.2 in 2017 (-6.6-10.6).

The rest of Table III shows that this pattern is robust to alternative testing approaches.[34] It holds in columns (2) and (3) when we estimate the slope for each year

[34]In these tests, we cluster standard errors by year and not by forecasted fiscal period, as we do in the rest of the paper, because observations are not available by forecasted fiscal period. Changing the level of clustering does not materially affect statistical inference.

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). Results from columns (3) and (5) are particularly remarkable. They show that the steepening of the term-structure is still present within industry (Column 3) and within analyst (Column 5), indicating that the aggregate trend is not driven by a change in sample composition by industry or analyst. To make sure that these results are not specific to the period 1983-1992 as our baseline, nor driven by I/B/E/S imperfect coverage at the beginning of the sample, we repeat the same analysis in Panel B excluding the 80s. The conclusion remains the same.

VI Social Media Data and Forecasts Informativeness

The informativeness of analysts' short-term forecasts has improved over time while that of long-term forecasts has declined. This pattern coincides with the rise of alternative data. Insofar as this reduces the cost of obtaining short-term information (our hypothesis), this evolution is therefore consistent with our main prediction (Corollary 1). Of course, many other mechanisms could explain the long-term evolution of the term-structure of forecast informativeness. For instance, investors may have become more short-term oriented, leading to analysts' compensation packages that reward the accuracy of short-term forecasts (an increase in γ in the model) and thus increase analysts' efforts to obtain short-term information (at the expense of their effort of obtaining long-term information). Moreover, corporate disclosure policies could have changed, for example, in response to regulatory changes that increase the penalty for disclosing imprecise information (e.g., via greater risk of litigation), creating a disincentive to share imprecise information (e.g., about long-term outcomes) with investors and analysts. Simultaneously, more stringent disclosure requirements may have induced firms to disclose more precise short-term information (e.g., Srinidhi et al. (2009)). More broadly, any factor resulting in a decrease in the cost of processing short-term information (parameter a) or an increase in the cost of processing long-term information (parameter b) could also explain our findings. Therefore, while in line with our theory, the long-term steepening of the term-structure of analysts' forecasts' informativeness is not direct evidence supporting it.

The ideal test of our hypothesis requires variation in analysts' access to short-term oriented alternative data (i.e., variation in a) that is orthogonal to other factors affecting

their effort allocation between the tasks of forecasting long-term and short-term earnings. To proxy for this, we exploit the introduction and progressive expansion of the financial social networking platform StockTwits (www.stocktwits.com) after 2009. We use this to construct measures of analysts' exposure to alternative (social media) data that, plausibly, (i) capture variations in the availability of alternative data used by analysts, and are (ii) unrelated to other forces driving the informativeness of their forecasts. To limit the scope for alternative explanations and isolate the economic channel underlying our theory, we further test a battery of ancillary predictions derived from the model.

A StockTwits Data

StockTwits was founded in 2008 as a social networking platform for investors to share their opinions about firms. Participants can post messages of up to 140 characters and 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.

We obtained data from StockTwits for all messages posted between January 1, 2009 and December 31, 2017. Similar to Cookson & Niessner (2020), for each message, we observe the user identifier, the date, content, and associated $cashtags with the corresponding tickers (a message can be associated with multiple tickers). We also observe specific information about both users and firms. We have access to users' self-declared information when they registered on the platform, including their name and investment horizon. For each firm discussed on StockTwits, we know its listing venue and its "watchlist", i.e., the number of users who explicitely follow that firm. For our analysis, we only keep messages about firms 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 VI about here]

Figure VI shows the evolution of the number of users and their posting intensity. 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 users on a firm's watchlist also increases sharply over time, up to about 2,000 in 2017. The lower panels display the evolution of the distributions of the daily numbers of messages and users on a firm's watchlists from 2009 to 2017. We note a substantial and increasing heterogeneity in the availability of StockTwit data across firms and time. This variation reflects the heterogeneous expansion of the platform over time, with some firms receiving high social media coverage early, some firms receiving coverage later, and others outside most discussions.

B Main Test Specification

Our test exploits StockTwits' heterogenous expansion across stocks to isolate the effect of social media data on analysts' forecasts. In this test, we compare how the informativeness of forecasts for a given horizon ($R^2_{i,t,h}$) changes after StockTwits' introduction for analysts with early and high exposure to StockTwits' data relative to analysts who were exposed later (or simultaneously but with less intensity). In essence, this approach is similar to a "difference-in-differences" approach in which variation in data generated on StockTwits about a given firm due to expanding coverage by StockTwits' users is used as a shock to analysts' exposure to alternative data (i.e., "treated" analyst are those covering that firm). We implement this methodology separately for different horizons to study the effect of alternative data on the overall term-structure of analysts' forecasts informativeness. Our test begins on January 1, 2005 – almost 5 years before the first message posted on the platform on July 13, 2009 – and ends on December 31, 2017. Our baseline specification by horizon h is:

$$R^2_{i,t,h} = \lambda(\text{Data Exposure})_{i,t-1} + \Gamma\text{Controls}_{i,t-1} + \eta_i + \eta_t + \omega_{i,t,h}, \tag{17}$$

where $R^2_{i,t,h}$ is the informativeness of the forecasts made by analyst i on day t for the forecasting horizon h. The parameters η_t and η_i are time and analyst fixed effects, controlling for common factors affecting the informativeness of all analysts, and for heterogeneous but time-invariant analyst-specific factors (observed and unobserved).[35] We further control

[35] All explanatory variables in eq.(17) are standardized by their sample standard deviation, are winsorized at the 1% and 99% by date t, and are measured at $t-1$. We cluster the standard errors of $\omega_{i,t}$ by forecasted fiscal period. Changing the level of clustering does not materially affect statistical inference.

for several characteristics of the firms covered by the analyst at $t-1$.[36]

The main variable of interest "Data Exposure" measures analyst i's exposure to social media data generated by StockTwits at $t-1$. It is equal to zero before we observe a message for the first time on the platform and then increases (differentially across analysts) with the expansion of StockTwits.[37] We posit that analysts who are more exposed to StockTwits data experience a positive shock to the volume of alternative social media data available to them. Thus, our hypothesis implies (see Corollary 1) that higher exposure leads to more informative short-term forecasts (i.e., $\lambda > 0$ for small h) but possibly less informative long-term forecasts (i.e., $\lambda < 0$ for large h).

As explained below, we measure the exposure of a given analyst to data generated by StockTwits in two distinct ways. One challenge is to ensure that these measures do not also capture how the analyst's exposure to data coming from *other* data sources changes after StockTwits' introduction. Messaging activity on StockTwits is indeed correlated with the arrival of information, whatever its origin, hence not all content on StockTwits necessarily originates from the discussions on the platform, but could come from (and relay information from) other sources, including those that provide access to traditional data (e.g., corporate news releases). Ideally, our measures should only capture the variation in a given analyst's exposure to data that is specifically generated by StockTwits and that would not (counterfactually) be available to her, without this social media site.

Our first measure of an analyst's exposure to StockTwits data relies on the number of users who have on their "watchlist" the same firms as the ones this analyst covers. A user's watchlist is a list of firms that the user follows. StockTwits aggregates this information at the firm level and, for each firm, reports the number of users having that firm on their watchlist. We aggregate this information at the analyst level by averaging

[36] Those characteristics include size (defined as log of total assets (inflation adjusted)), age (defined as log of age since public listing)), cash-flow to assets ratio, cash to assets ratio, debt to assets ratio, and Tobin's Q. All variables are averaged daily across firms covered by analyst i on day t. All accounting-based variables are computed using the last available annual accounting information. Tobin's Q is computed using last stock price information. Detailed definitions for each characteristic are provided in the appendix.

[37] Thus, the variable measuring "treatment" here is continuous and not dichotomous as is usually the case in a standard difference-in-differences. In a standard setting, the treatment effect is typically estimated from a binary variable incremented by one after treated units receive the treatment, controlling for unit and time fixed effects (as we also do). This estimation strategy is not feasible here because discussions on StockTwits did not reach a steady state in a short time.

this number across the firms she covers. We then use this average number of users (denoted *#Watchlist*) as a measure of exposure to StockTwits data.[38] Importantly, a user's watchlist is persistent. A user typically declares it when she registers on StockTwits, and rarely modifies it. As a result, a firm's watchlist changes because new users register and enter the platform. Therefore, the source of variation in *#Watchlist* mostly reflects the overall expansion of StockTwits, both over time and across firms, and not the arrival of information from other sources. In fact, as shown below, changes in a given firm's watchlist are largely uncorrelated with the arrival of information from traditional data sources (which could have affected analysts' forecasts in the absence of StockTwits).

Our second measure relies on the number of messages posted about the firms covered by an analyst. However, we do not consider the number of actual messages. Instead, we use the number of *hypothetical* messages, which we estimate at the firm level, and then aggregate at the analyst level. Specifically, for each firm, we compute its share of the total number of messages exchanged on StockTwits since inception until day $t-1$. We next define the number of hypothetical messages on day t for a given firm j as the total number of messages on Stocktwits on this day multiplied by firm j's historical share of messages on StockTwits on day $t-1$. Intuitively, this variable measures the number of messages that one would expect to observe on day t if the intensity of discussions about firm j relative to the intensity of discussions about other firms is at its average historical level. While changes in the number of actual messages are correlated with the concurrent arrival of information from traditional data sources, changes in the hypothetical number of messages are not (as we show below). Finally, we compute the total number of hypothetical messages in the last thirty days (from $t-30$ to $t-1$) for all firms, and aggregate at the analyst level by taking the average across the firms she covers (denoted *#Hypothetical Messages*).

A common and key feature of *#Watchlist* and *#Hypothetical Messages* is the persistence in their *relative* cross-sectional heterogeneity. The share of total users that a firm captures in the variable "watchlist" quickly becomes stable after StockTwits introduction, such that the *relative* number of "watchers" a firm has compared to other firms is

[38] The coverage of firms by StockTwits increases progressively over time (see Figure VI). Thus, if a firm covered by an analyst is not yet covered by StockTwits on a given day in our sample, we set its number of users to zero.

(mostly) time invariant. Likewise, the market share of all StockTwits messages varies between firms but is stable within firm, so that the *relative* level of discussion intensity across firms is (almost) constant. This high degree of persistence is important because it implies that most of that relative heterogeneity across firms will be controlled for in our specification.[39] As a result, the main source of variation for our measures is either the *aggregate* number of users (*#Watchlist*) or the *aggregate* number of messages (*#Hypothetical Messages*), which we believe are both plausibly unrelated to individual firm and analyst characteristics, as well as the regular flow of firm-level information (as shown later).

[Insert Table IV about here]

Table IV presents summary statistics of the sample used for our main test. This sample contains 30,958,705 observations. The average analysts' 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 to StockTwits 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 hypothetical messages for firms covered by the average analyst is equal to 13 with a standard variation of 43.

C Is StockTwits a Good Laboratory for our Test?

Before presenting the results, it is worth discussing why the estimation of eq.(17) is a good test of the economic mechanism highlighted in our model. The first reason is that, as explained in the previous section, both measures of analysts' data exposure are built to be unrelated to information that would be available to analysts even in the absence StockTwits (e.g., from firms' disclosures or news platforms). To support this claim, we test whether the number of users in a given firm's watchlist and the number

[39] It is differenced out by the analyst fixed effects in eq.(17), assuming analysts always cover the same firms. To make sure that our results are not driven by a violation of this assumption, we verify and show in Table A.5 that our main results survive when focusing on a sub-sample of analysts covering always the same firms.

of hypothetical messages about it correlate with the arrival of traditional information, as measured by the number of distinct news about that firm recorded in Capital IQ - Key Development. This dataset monitors more than 230 categories of news (e.g., executive changes, announcements of M&A, earnings, dividend, delayed filings, or SEC inquiries) and includes almost 12 million news related to firms in our sample.[40] Table A.2, reported in the Appendix, confirms the absence of any significant relationship between the number of recent news items (measured over different periods) and (a) the number of users (in a firm's watchlist) or (b) hypothetical messages.

Second, the existing literature has shown that the information generated on social media is primarily relevant for forecasting firms' short-term returns, sales, and earnings.[41] Moreover, Figure VII shows that the vast majority of StockTwits' messages come from users that are either "day traders" (35.4%) or "swing traders" (49%).[42] Only a small fraction of messages are issued by users declaring long horizons, either "position traders" (6.2%) or "long-term investors" (8.6%).

[Insert Figure VII about here]

Last, it is likely that analysts use social media platforms like StockTwits as a source of information. Indeed, several industry reports highlight the potential of social media as sources of information for investors.[43] Consistent with this interest, StockTwits' datafeed has been gradually integrated into all major financial information aggregation platforms commonly used by practitioners to source information about firms and industries (e.g., Bloomberg.com or Reuters.com, among others), making it likely that analysts are exposed to data generated on StockTwits. We further show in the Appendix (see Table A.3) that analysts are more likely to issue (or revise) a forecast on a given firm and day following an increase in StockTwits' messaging intensity in the prior thirty days (even when there is no information from traditional data sources over that period). In addition, using biographic

[40] We consider all news except M&A rumors, because those rumors may come from social media outlets.

[41] See, for instance, Chen et al. (2014), Jame et al. (2016), Renault (2017), or Bartov et al. (2020).

[42] To build this figure, we use the fact that StockTwits users can self-declare their investment horizon as being in one of four categories: "day trader", "swing trader", "position trader", and "long-term investors." 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.

[43] See, for instance, Deustche Bank Market Research, March 4 2014; "Demystifying Alternative Data", Greenwhich Associates 2019; or "Big Data & Investment Management", Citi Business Advisory Services, 2015).

information (analysts' last names and the first letter of their first names) from I/B/E/S (obtained from the price target dataset between 2009 and 2017), we find that 35% of (7,656 distinct) analysts' names exactly match that of StockTwits' account holders. An account is not required to consume information from StockTwits but is necessary to receive alerts when new messages are posted about a firm. Thus, analysts may register on StockTwits to get automatic alerts when a firm they cover is discussed by users.[44]

D Main Results

Table V presents estimates of eq.(17) separately across four distinct groups of horizons ranging from one year or less ($h \leq 1$) to more than three years ($h \geq 3$), using our two measures of "Data Exposure" (*#Watchlist* in Panel A and *#Hypothetical Messages* in Panel B).[45] To facilitate economic interpretation, we normalize both variables by their sample standard deviation.

[Insert Table V about here]

Columns (1) and (2) of Table V show that increased exposure to StockTwits' data has a significantly *positive* effect on the informativeness of analysts' short-term forecasts ($h \leq 1$; horizon less than one year). Columns (3) and (4) indicate that increased exposure to StockTwits does not significantly affect the informativeness of forecasts at mid-term horizons ($1 < h \leq 2$). In contrast, Columns (5) to (8) show that increased exposure to StockTwits has a significantly *negative* influence on the informativeness of long-term forecasts ($2 < h \leq 3$ or $h \geq 3$). A one standard deviation increase in analysts' exposure to StockTwits' data leads to a drop in the informativeness of analysts' long-term forecasts of 1.51% to 1.88%, and an improvement in the informativeness of their short-term forecasts of 0.53% and 0.78%. These results are in line with the model's predictions.

To provide a different perspective on the economic magnitude of these effects, we modify our baseline specification (eq.(17)) by pooling analyst-day-horizon observations across all horizons, and include an interaction term between "Data Exposure" and the

[44] Our mechanism and tests do not require analysts to actually post information on StockTwits. However, we note that they actually do so on "Estimize", another social media specialized in earnings' forecasting. See "Estimize: Crowdsourced financial estimates and data" (on Estimize website) and Jame et al. (2016).

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

(annualized) forecasting horizon of each observation (centered at a one-year horizon for convenience). More specifically, we estimate:

$$R^2_{i,t,h} = \lambda_0(\text{Data Exposure})_{i,t-1} \times (h-1) + \lambda_1(h-1) + \lambda_2(\text{Data Exposure})_{i,t-1} + ... + \omega_{i,t,h}. \tag{18}$$

Estimates of this specification are reported in Table VI. Columns (1) and (4) reveal that the coefficients on the interaction term (λ_0) are negative and statistically significant. Thus, greater exposure to data generated by StockTwits steepens the term-structure of analysts' forecast informativeness. Column (1) (respectively column (2)) indicates that, for a given increase in data exposure, 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 forecast informativeness by 16.66% (16.64%) for analysts that are not exposed to StockTwits' data. This decline rises by 0.86% (0.56%) for each standard deviation increase in analysts' 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 analysts' forecase informativeness within a given annual forecasting horizon (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 forecasts issued on a given day and for a given horizon.

[Insert Table VI about here]

E Additional Predictions and Ancillary Results

To further support the mechanism highlighted by our model, we test three ancillary predictions. Namely, the steepening of the term-structure of analysts' forecasts informativeness resulting from the increased availability of alternative data should be more pronounced when (i) such data contains more short-term information, so that it is more likely that access to these data reduces the marginal cost of producing short-term information (a), (ii) the cost of multi-tasking (c) is high, and (iii) firms' earnings are less auto-correlated (so that the condition $\beta < (\frac{c}{2b})^{\frac{1}{2}} \frac{\xi_{lt}}{\xi_{st}}$ in Corollary 1 is more likely to be satisfied). We find broad support for these specific predictions.

E.1 Users' Investing Horizon (a)

To assess whether the effect of exposure to StockTwits' data on analysts' forecast informativeness is stronger when that data are more likely to generate short-term information, we exploit the heterogeneity in investing horizon declared by StockTwits' users and posit that users who define themselves as "day traders" are more likely to collectively produce information about the short-term than those who define themselves as "long-term investors". We count the number of hypothetical messages posted over the last thirty days by each category of users for each firm and compute the average number of hypothetical messages for each category across firms covered by each analyst (and day). We then reestimate eq.(18) with a separate interaction term for the average number of hypothetical messages of each category. Table VII reports the results. In line with our mechanism, the steepening of the term-structure of forecasts' informativeness is stronger for analysts exposed to data generated by StockTwits' users with short-term horizons. In fact, the interaction terms between hypothetical messages and forecasting horizon is significantly negative only for messages posted by "day traders" (coefficients ranging between -0.49 and -0.88). We detect no significant effects for the three other categories of users.

[Insert Table VII about here]

E.2 Multi-Tasking Costs (c)

In our theory, a decrease in the marginal cost of producing the short-term signal steepens the term-structure of analysts' forecasts informativeness if the cost of multi-tasking is high enough. Thus, the steepening of the term-structure of forecast informativeness should be stronger for analysts facing higher costs of multi-tasking. This is likely to be the case for those who cover more stocks, since the total number of forecasting tasks (within and across firms) increases with coverage. For instance, Harford et al. (2019) state (p.2182) that "busy" analysts (those covering larger portfolios) are "*more likely to hit the constraint created by analysts' limited time, energy, and resources, making it even more critical to be strategic in their research activities*" and provide evidence that this is the case (see their Table 6). Results in Table VI confirm this prediction. We re-estimate the specifications reported in Table VI interacting our measures of "Data Exposure" with the number of firms in analysts' portfolio. Consistent with our conjecture, coefficients on the triple

interaction term between data exposure, horizon, and the number of covered firms are all negative, and five out of six are statistically significant.

[Insert Table VIII about here]

E.3 Correlated Earnings (β)

Our model also predicts that a drop in the marginal cost of producing the short-term signal should lead to a stronger steepening of the term-structure of forecasts informativeness when the correlation between the long and short-term earnings of the covered firms (a proxy for β in the model) is smaller. We test this prediction using firms' earnings auto-correlation as an empirical proxy for β for each analyst (and day), obtained by regressing firms' quarterly earnings on their lag (without a constant) using a rolling window of two years (and requiring at least four observations), and averaging the estimated auto-correlation across firms covered by the analyst. As predicted, Table IX indicates that the coefficients on the triple interactions between "Data Exposue", horizon, and earnings auto-correlations are all positive and statistically significant. That is, the negative association between analysts' exposure to data generated by StockTwits and the informativeness of their long-term forecasts is less pronounced for analysts covering firms whose earnings are *more* auto-correlated.

[Insert Table IX about here]

F Alternative Explanations and Intepretations

Our empirical findings are consistent with our prediction: an increase in analysts' exposure to StockTwits' data increases the informativeness of their short-term forecast but decreases that of their long-term forecasts. Moreover, the heterogeneity of this effect across analysts can be explained by our theory. One legitimate concern is however that our measures of analysts' data exposure may co-vary with unobserved variables affecting the informativeness of analysts' forecasts through other channels. If this is the case, the effects that we attribute to a change in the volume of alternative data available to analysts (due to the introduction of StockTwits) may in fact stem from other factors. In our context, potential omitted variables can be broadly classified into three categories.

The first category includes variables related to the amount of publicly available information. Indeed, forecasts are more informative when more public information is available. Therefore, one concern (as discussed above) is that our measures of analysts' exposure to StockTwits' data correlate with other sources of information available to them. In that case, our findings could be due to changes in information available to analysts from traditional sources rather than an expansion in analysts' sources of information due to the introduction of StockTwits. However, as explained in Sections VI.B and VI.C, our measures of analysts' data exposure are specifically built to avoid this problem, and in fact, Table A.2. suggests that these measures do not covary with the volumne of public news identified by Capital IQ-Key Development. According to WRDS, Key Development tracks all material news and events potentially affecting the market value of a firm from various sources (including press releases, news wires, regulatory filings, exchanges, company web sites, and call transcripts) with significant coverage starting from 2003 (more than five years before the start of StockTwits' activity). Thus, we believe that it is unlikely that Key Development systematically misses relevant news. To further mitigate this concern, we control in our main tests for the average trading volume of the firms covered by analysts over the last 30 days. In this way, we control for the potential effect of news (public or private) that are material enough to generate trading. Table A.4 reported in Appendix shows that controlling for trading volume does not change our results.

The second category includes variables related to earnings uncertainty. When uncertainty is higher, forecasts are naturally less precise, and thus less informative, so another concern is that our results arise because uncertainty about the earnings of the covered firms changes over the 2005-2017 period. Short(long)-term earnings may have become less (more) uncertain, and our measures of data exposure may systematically reflect these changes (e.g., users including firms with less uncertain short-term earnings in their watchlist). However, as discussed in Section VI.C, our measure of forecasts informativeness is explicitly designed to isolate forecasting informativeness from forecasting difficulty (absolute informativeness – the numerator in eq.(16) – is normalized by total uncertainty – the denominator).

Changes in earnings uncertainty during the period 2005-2017 may still indirectly affect forecasts' informativeness by altering analysts' incentives to process information. More

generally, the third and last category of possible omitted variables includes all variables affecting how analysts allocate their effort between the tasks of forecasting short-term and long-term earnings, including compensation schemes, career prospects, investors' demand, or brokers' internal organization. Nevertheless, to explain our results, changes in these variables should trigger a simultaneous increase in the informativeness of short-term forecasts and a decrease in that of long-term ones. Moreover, they should systematically coincide with the timing of StockTwits expansion across firms, as well as the aggregate variation in messaging activity on StockTwits, while being completely unrelated to StockTwits (i.e., the same changes in analysts' incentives at the exact same time would have been observed, absent StockTwits). We cannot rule out this scenario, but it seems unlikely.

Another possibility is that the introduction of StockTwits *directly* influences analysts' allocation of effort between short-term and long-term forecasting via another channel than that highlighted in our model (i.e., a change in the relative cost of producing short-term and long-term signals, combined with multi-tasking costs). For instance, StockTwits may help analysts to learn about investors' demand for short-term and long-term information and induce them to allocate effort accordingly. In this scenario, our results would still be attributed to the existence of StockTwits but the underlying economic mechanism would be different. We note that the stronger effect of StockTwits on forecasts' informativeness when multi-tasking costs are high and when earnings auto-correlation is low is consistent with the channel we propose. Although we cannot exclude that StockTwits influences analysts' allocation of effort via other channels, we reason that other channels (e.g., the investors' demand channel) would induce analysts to modify their coverage (e.g., by tilting their coverage towards stocks with more investors' interest). Table A.5 in the Appendix shows however that our results remain broadly unchanged when we focus on analysts who always cover the same firms (representing 46% of our sample), for which incentives are likely stable.

VII Conclusion

This paper examines how alternative data affect the informativeness of financial forecasts at various horizons. We posit that alternative data reduce the cost of producing information about short-term cash-flows relatively more than that about long-term cash-flows.

We show theoretically that this shift can induce forecasters to focus more on the production of short-term information, at the expense of the informativeness of forecasts of long-term cash-flows. Our main contribution is to test this novel prediction and confirm it. Specifically, we find empirically that an expansionary shock to available alternative data (namely, the introduction and gradual expansion of StockTwits) is associated with a drop in the informativeness of sell-side equity analysts' forecasts of long-term (more than two years) earnings, even though the informativeness of their short-term (less than one year) forecasts improves. This finding suggests that the emergence of alternative data could negatively affect the long-term informativeness of asset prices and the efficiency of investment decisions. Studying these questions are interesting avenues for future research.

References

Abis, S. (2018), Man vs machine: Quantitative and discretionary equity management, Technical report.

Bai, J., Philippon, T. & Savov, A. (2016), 'Have financial markets become more informative?', *Journal of Financial Economics* **122**, 625–654.

Bandyopadhyay, S., Brown, L. & Richardson, G. (1995), 'Analysts' use of earnings forecasts in predicting stock returns: Forecast horizon effects', *International Journal of Forecasting* (11), 429–445.

Bartov, E., Faurel, L. & Mohanram, P. (2020), Can twitter help predict firm-level earnings and stock returns? Working Paper.

Begeneau, J., Farboodi, M. & Veldkamp, L. (2018), 'Big data in finance and the growth of large firms', *Journal of Monetary Economics* **97**(1), 71–87.

Chen, H., De, P., Hu, Y. & Hwang, B.-H. (2014), 'Wisdom of crowds: The value of stock opinions transmitted through social media', *Review of Financial Studies* (27), 1367–1403.

Cookson, A. & Niessner, M. (2020), 'Why don't we agree? evidence from a social network of investors', *Journal of Finance* **75**(1), 173–228.

Cookson, T., Engelberg, J. & Mullins, W. (2020*a*), Does partisanship shape investor beliefs? evidence from the covid-19 pandemic. Review of Asset Pricing Studies (forthcoming).

Cookson, T., Engelberg, J. & Mullins, W. (2020*b*), Echo chambers. Working Paper.

Crane, A. & Crotty, K. (2020), 'How skilled are security analysts?', *Journal of Finance* **75**(3), 1629–1675.

Davila, E. & Parlatore, C. (2020), Identifying price informativeness. Working paper.

Dugast, J. & Foucault, T. (2018), 'Data abundance and asset price informativeness', *Journal of Financial Economics* (130), 367–391.

Farboodi, M., Matray, A., Veldkamp, L. & Venkateswaran, V. (2020), Where has all the big data gone? Working Paper.

Farboodi, M. & Veldkamp, L. (2020), 'Long rung growth of financial data technology', *Forthcoming in the American Economic Review* .

Froot, K., Kang, N., Ozik, G. & Sadka, R. (2017), 'What do measures of real-time corporate sales say about earnings surprises and post-announcement returns?', *Journal of Financial Economics* (125), 143–162.

Gao, M. & Huang, J. (2020), Informing the market: The effect of modern information technologies on information production. Review of Financial Studies (forthcoming).

Giannini, R., Irvine, P. & Shu, T. (2019), 'The convergence and divergence of investors' opinions around earnings news: Evidence from a social network', *Journal of Financial Markets* (42), 94–120.

Goldfarb, A. & Tucker, C. (2019), 'Digital economics', *Journal of Economic Literature* **57**(1), 3–43.

Goldstein, I. & Yang, L. (2015), 'Information diversity and complementarities in trading and information acquisition', *Journal of Finance* **70**, 1723–17–5.

Goldstein, I., Yang, S. & Zuo, L. (2020), 'The real effects of modern information technologies', *Working paper, NBER* .

Grennan, J. & Michaely, R. (2019), 'Fintechs and the market for financial analysis', *Forthcoming Journal of Financial and Quantitative Aanalysis* .

Grennan, J. & Michaely, R. (2020), Artificial intelligence and the future of work: Evidence from analysts. working paper.

Grossman, S. & Stiglitz, J. (1980), 'On the impossibility of informationally ef?cient markets', *American Economic Review* (70), 393–408.

Harford, J., Jiang, F., Wang, R. & Xie, F. (2019), 'Analyst career concerns, effort allocation, and firms' informational environment', *Review of Financial Studies* **32**(6), 2179–2224.

Hilary, G. & Hsu, C. (2013), 'Analyst forecast consistency', *Journal of Finance* (68), 271–297.

Hirshleifer, D., Levi, Y., Lourie, B. & Teoh, S. H. (2019), 'Decision fatigue and heuristic analyst forecasts', *Journal of Financial Economics* pp. 83–98.

Hong, H. & Kacperczyk, M. (2010), 'Competition and bias', *Quarterly Journal of Economics* **125**, 1683–1725.

Huang, S., Xiong, Y. & Yang, L. (2020), 'Skills acquisition and data sales', *Working paper* .

Jame, R., Johnston, R., Markov, S. & Wolfe, M. (2016), 'The value of crowdsourced earnings forecasts', *Journal of Accounting Research* **54**, 1077–1109.

Katona, Z., Painter, M., Patatoukas, P. N. & Zeng, J. (2019), On the capital market consequences of alternative data: Evidence from outer space. Working Paper.

Merkley, K., Michaely, R. & Pacelli, J. (2017), 'Does the scope of the sell-side analyst industry matter? an examination of bias, accuracy, and information content of analyst reports', *Journal of Finance* **72**(4), 653–686.

Mest, D. P. & Plummer, E. (1999), 'ransitory and persistent earnings components as reflected in analysts' short-term and long-term earnings forecasts: evidence from a nonlinear model', *International Journal of Forecasting* **15**, 291–308.

Monsell, S. (2003), 'Task switching', *Trends in Cognitive Science* **7**(3), 134–140.

Myatts, D. P. & Wallace, C. (2012), 'Endogenous information acquisition in coordination games', *Review of Economic Studies* pp. 340–374.

Renault, T. (2017), 'Intraday online investor sentiment and return patterns in the u.s. stock market.', *Journal of Banking and Finance* **84**, 25–40.

Srinidhi, B., Leung, S. & Jaggi, B. (2009), 'Differential effects of regulation fd on short- and long-term analyst forecasts', *Journal of Accounting and Public Policy* **28**, 401–418.

van Binsbergen, J. H., Han, X. & Lopez-Lira, A. (2020), 'Man vs. machine learning: The term structure of earbnings expectations and conditional biases', *Working paper, NBER* .

Verrecchia, R. (1982), 'Information acquisition in a noisy rational expectations economy', *Econometrica* pp. 1415–1430.

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

Zhu, C. (2019), 'Big data as a governance mechanism', *Review of Financial Studies* **32**(5), 2021–2061.

Figure I: Timeline of the model

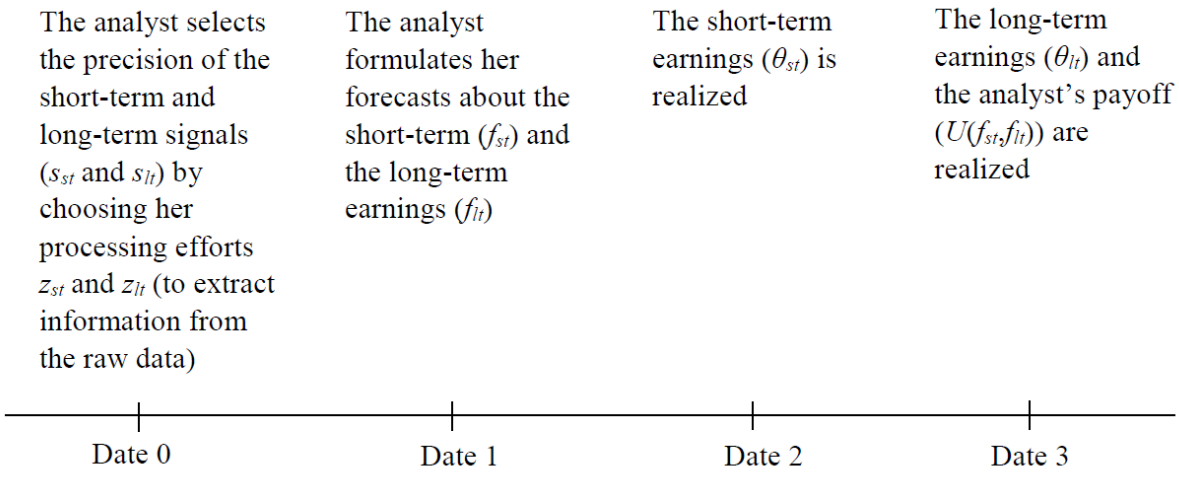
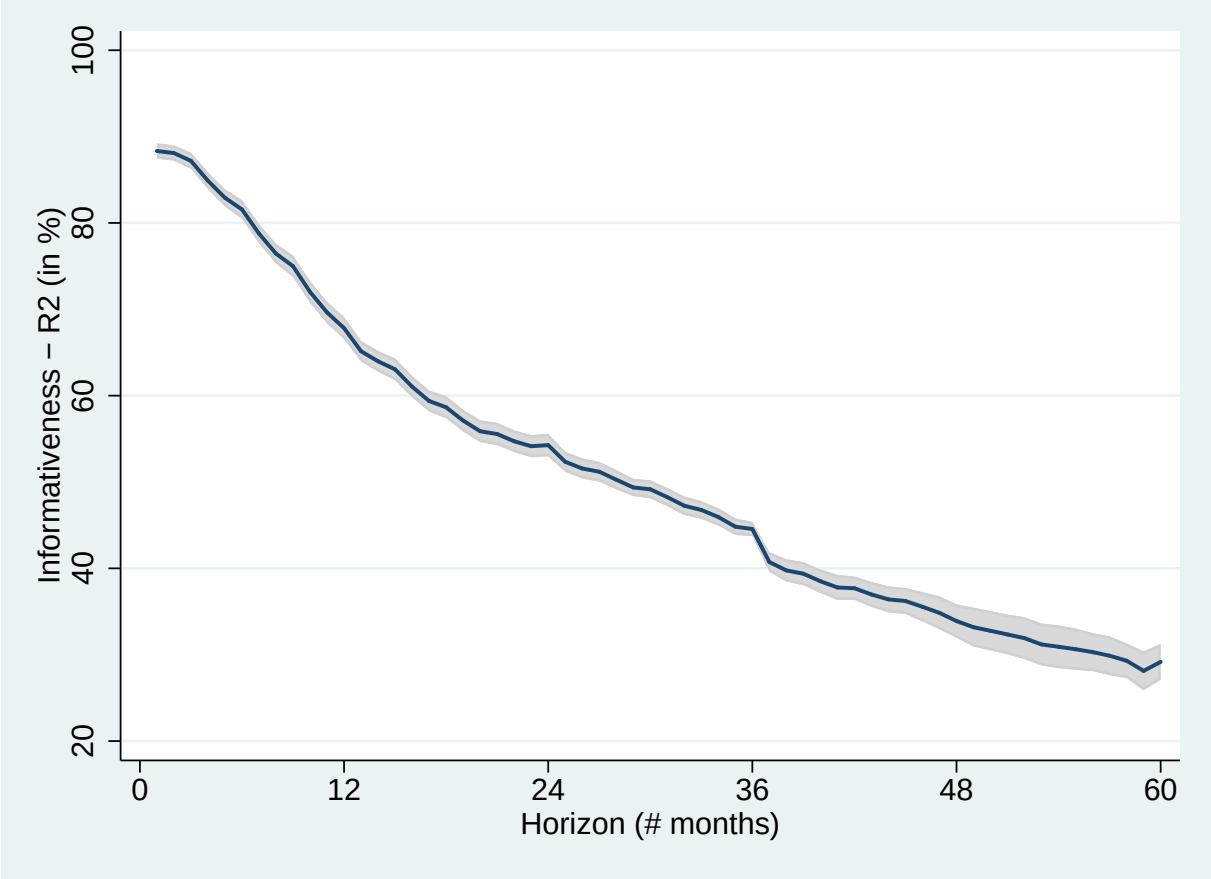
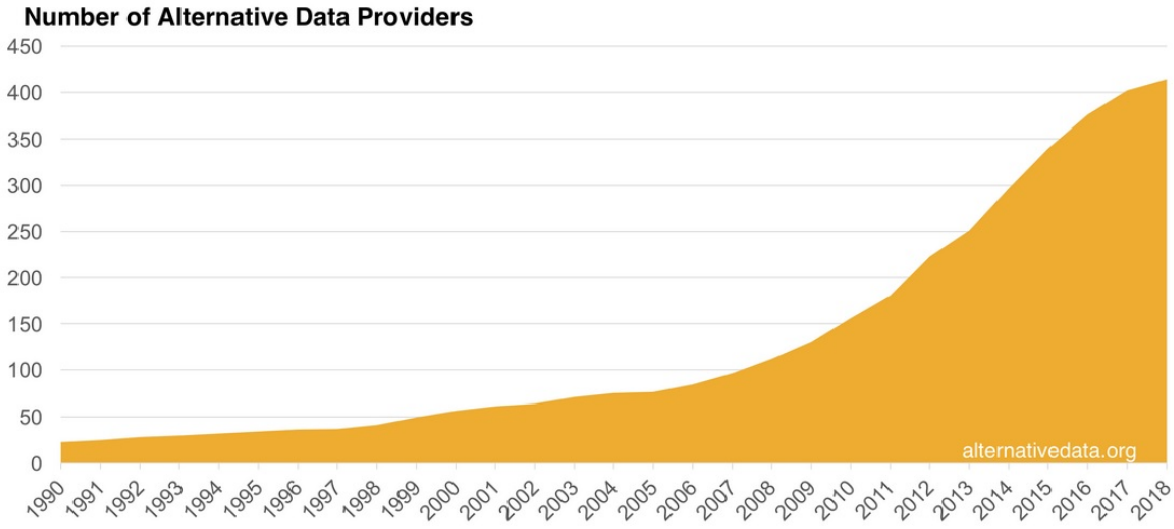


Figure II: The term-structure of analysts forecasts' informativeness



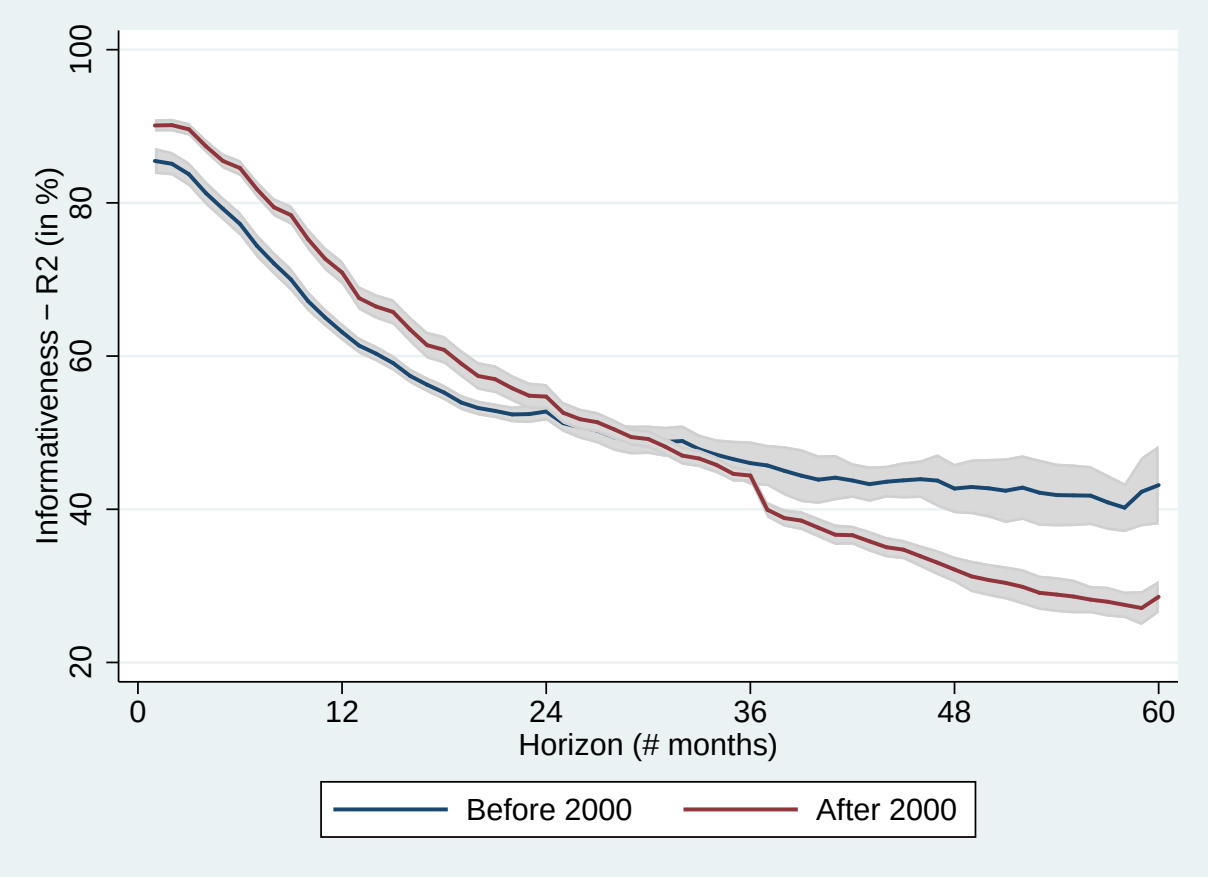
This figure displays the term-structure of analysts forecasts' informativeness. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in months) from zero to five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



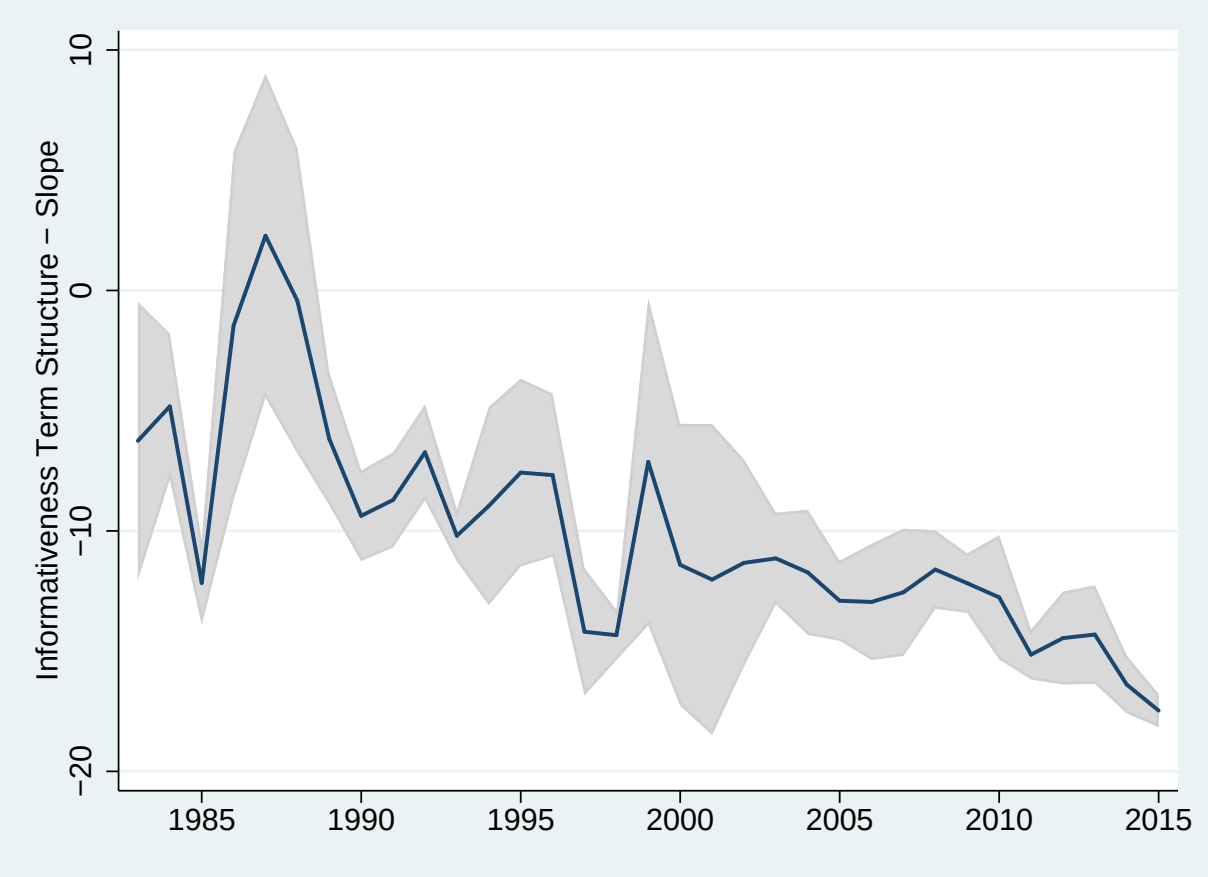
This figure displays the evolution of the number of alternative data providers reported by the website alternativedata.org. The graph was taken from https://alternativedata.org/stats/ (on 12/23/2020).

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



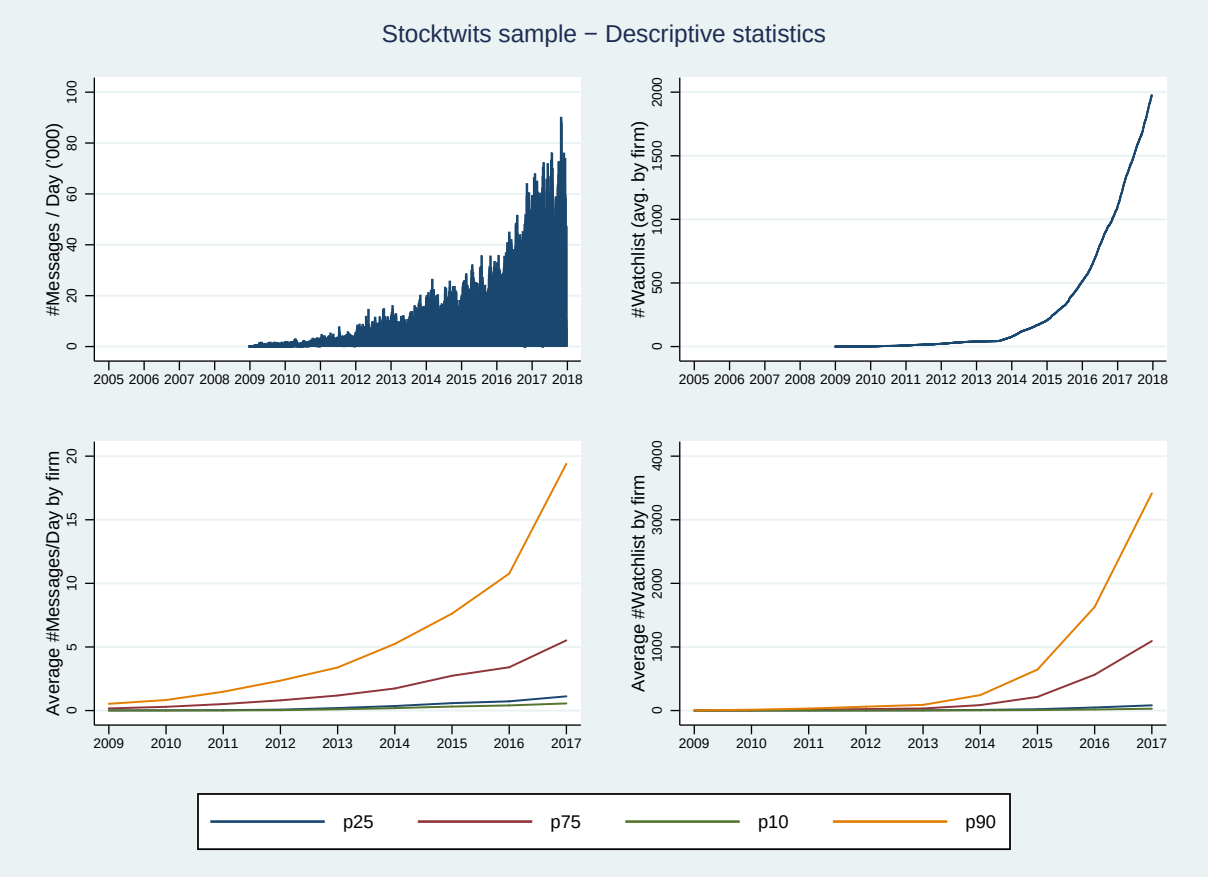
This figure displays the term-structure of analysts forecasts' informativeness before and after 2000. It is obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on a set of horizon binary variables measuring all possible horizons (in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



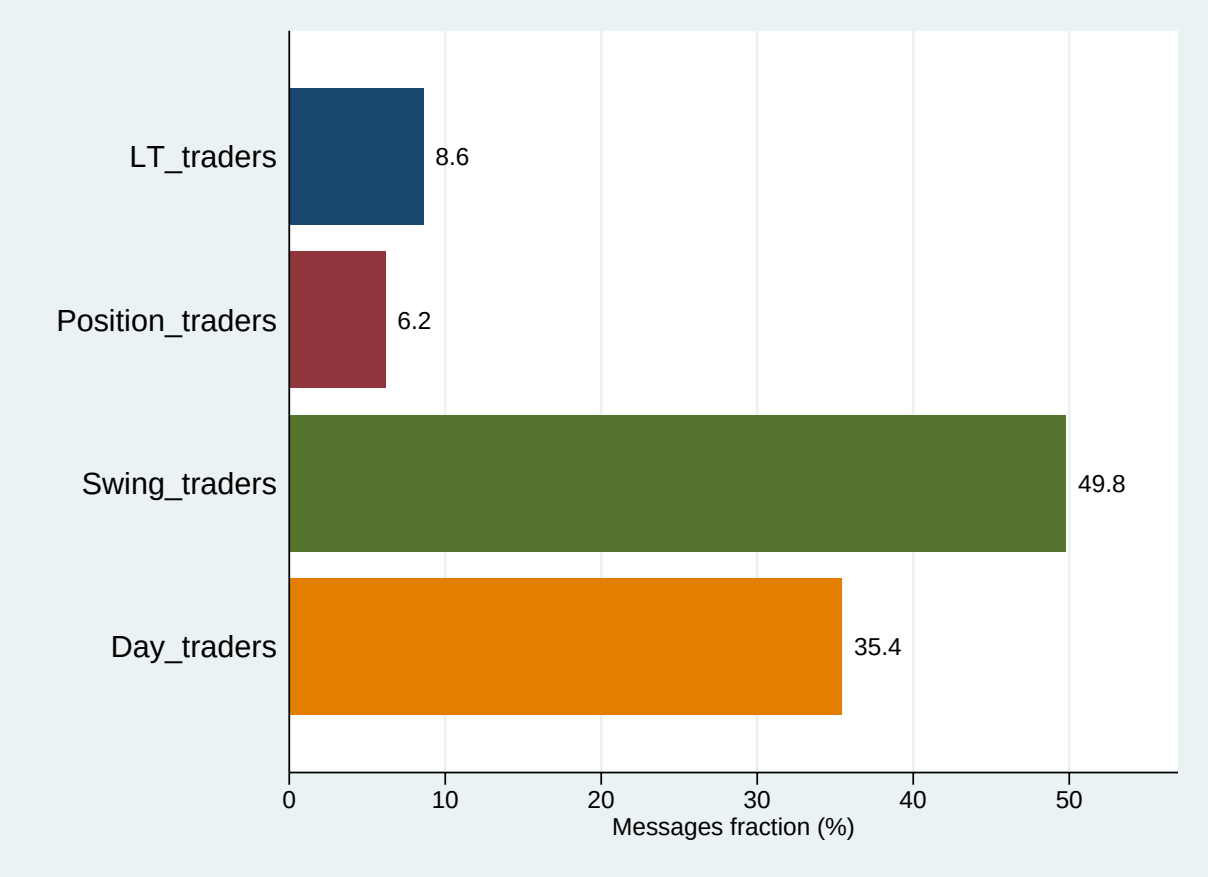
This figure displays the evolution of the slope of the term-structure of analysts forecasts' informativeness. The annual slopes are obtained by regressing the informativeness of the forecasts made by an analyst on a given day for a given horizon (R^2) on horizon (measured as the number of days between the forecasting date and the date of actual earnings release divided by 365), separately for every calendar year. The figure plots the resulting slope coefficients. Each slope coefficient measures how informativeness changes for every annual increment of horizon. For example, a slope coefficient of -11 in 2000 indicates that in 2000 (R^2) decreases on average by 11 percentage points when forecasting horizon increases by 1 year. The shaded gray area corresponds to a 90% confidence interval.

Figure VI: 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 number of users that have a given firm in their watchlist (averaged across firms). A user's watchlist is a list of firms that the user follows. StockTwits aggregates this information at the firm level and reports the number of users having that firm on their watchlist. The graph shows how this number has changed over time across firms. In our main tests, we further aggregate this information at the analyst level by averaging this number across the firms covered by the analyst. 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 VII: 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 investor". 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 (Table II and Table III). 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. #Firms is the number of firms covered by an analyst on a forecasting day used to compute R^2. 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
#Firms	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
#Firms	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
#Firms	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
#Firms	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
#Firms	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
#Firms	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. Some specifications include fixed effects for 2-digit SIC industry (the main industry of analysts' portfolio), firms' size quintiles and age (based on average firm size and firm 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. #Firms is the number of firms 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. It is set to zero prior to StockTwits' introduction in 2009. #Messages and #Hypothetical Messages is the average number of actual and hypothetical messages posted about firms (in the last thirty days) that analyst covers on a given day. Both variables are set to zero prior to StockTwits' introduction in 2009. Auto-correlation is the average earnings' autocorrelation across the firms that an analyst covers on a given day. The other variables are control variables used in the analysis. Detailed variable definitions are provided 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
#Firms	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
#Hypothetical Messages	30,958,706	13	43	0	0	2	9	990
Auto-correlation	29,364,398	0.67	0.21	-0.01	0.55	0.69	0.82	1.12
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

Table V: Data exposure and forecasts informativeness by horizon

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to analysts' exposure to social media data generated by StockTwits (eq.(17). The total sample includes all available analyst-day-horizon observations between 2005 and 2017, which we split by forecasting horizon sub-sample. 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. Data Exposure is an aggregate measure of an analyst's exposure to data generated on StockTwits, 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. Data exposure is set to zero prior to StockTwits' introduction in 2009, and further normalized by its in-sample standard deviation. In panel A, Data Exposure is based on the number of users that have the firm in their watchlist. In Panel B, Data Exposure is based on the number of hypothetical messages posted about a firm from $t-30$ to $t-1$. 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 Data Exposure based on #Watchlist								
Data Exposure	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 Data Exposure based on #Hypothetical Messages								
Data Exposure	0.78***	0.69***	0.76	0.30	0.25	-0.99***	-1.88***	-1.87***
	(4.77)	(4.31)	(1.53)	(0.64)	(-0.53)	(-2.73)	(-4.25)	(-3.52)
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: Data exposure 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 generated by StockTwits (eq.(18). The sample includes all available analyst-day-horizon observations between 2005 and 2017. 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. Data Exposure is an aggregate measure of analysts' exposure to data generated by StockTwits, 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. Data Exposure is set to zero prior to StockTwits introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the number of users that have the firm in their watchlist, or the number of hypothetical messages posted 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 Data Exposure can be interpreted as the unconditional effect on one-year informativeness. 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)
Data Exposure:	#Watchlist			# Hypothetical Messages		
Horizon × Data Exposure	-0.86***	-0.78***	-0.96***	-0.56**	-0.81***	-0.89***
	(-2.59)	(-3.06)	(-3.72)	(-2.32)	(-4.06)	(-4.26)
Data Exposure	0.13	-0.17	-0.35	0.58**	0.07	-0.15
	(0.50)	(-0.64)	(-1.29)	(2.11)	(0.25)	(-0.55)
Horizon	-16.66***			-16.54***		
	(-33.86)			(-32.16)		
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 generated on StockTwits. The sample includes all available analyst-day-horizon observations between 2005 and 2017. 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 report their usual investing horizon by declaring themselves as "Day Traders", "Swing Traders","Position Traders", or "Long-term investors". In columns (1) to (3), we measure an analyst's exposure to data produced by each category of users using the number of hypothetical messages posted by these users. For each category, we count the total number of hypothetical messages posted from $t-30$ to $t-1$ about the firms covered by the analyst, and then divide this number by the number of firms the analyst covers. 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 coefficients on the variables Data Exposure can be interpreted as the unconditional effect on one-year informativeness. All measures of Data Exposure are set to zero prior to Stocktwits introduction in 2009, and are further normalized by their in-sample standard deviation. 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)		
Data Exposure:	# Hypothetical Messages		
OLS	(1)	(2)	(3)
Horizon × (#Hypothetical Messages by Day Traders)	-0.88*	-0.49**	-0.49**
	(-1.66)	(-2.05)	(-2.11)
Horizon × (#Hypothetical Messages by Swing Traders)	0.06	-0.62	-0.62
	(0.08)	(-1.44)	(-1.53)
Horizon × (#Hypothetical Messages by Position Traders)	0.22	0.36	0.38
	(0.71)	(1.36)	(1.48)
Horizon × (#Hypothetical Messages by LT Traders)	-0.23	-0.05	-0.12
	(-1.29)	(-0.26)	(-0.67)
#Hypothetical Messages by Day Traders	0.64	0.10	0.21
	(1.17)	(0.20)	(0.41)
#Hypothetical Messages by Swing Traders	-0.40	0.03	-0.30
	(-0.81)	(0.05)	(-0.53)
#Hypothetical Messages by Position Traders	0.85***	0.53	0.41
	(3.37)	(1.55)	(1.30)
#Hypothetical Messages by LT Traders	-0.23	-0.35	-0.32
	(-0.71)	(-1.12)	(-1.01)
Horizon	-16.60***		
	(-30.55)		
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 generated by StockTwits. The sample includes all available analyst-day-horizon observations between 2005 and 2017. 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. Data Exposure is an aggregate measure of analysts' exposure to data generated by StockTwits, 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. Data Exposure is set to zero prior to StockTwits introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based either on the number of users that have the firm in their watchlist, or the number of hypothetical messages posted 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 Data exposure can be interpreted as the unconditional effect on one-year informativeness. #Firms is the number of firms covered by an analyst on a given forecasting day. 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)					
Data Exposure:	#Watchlist			# Hypothetical Messages		
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Data Exposure × #Firms	-0.14***	-0.06***	-0.06***	-0.10***	-0.05*	-0.06***
	(-5.71)	(-3.38)	(-3.82)	(-6.81)	(-1.94)	(-2.65)
Horizon × Data Exposure	0.69	-0.04	-0.23	-0.05**	-0.02	-0.02
	(1.61)	(-0.10)	(-0.74)	(-1.97)	(-0.81)	(-0.92)
Horizon × #Firms	-0.15***	-0.23***	-0.23***	-0.14***	-0.23***	-0.22***
	(-6.58)	(-8.67)	(-8.24)	(-5.86)	(-8.41)	(-7.86)
Data Exposure × #Firms	-0.09***	-0.05***	-0.04**	-0.05**	-0.02	-0.02
	(-3.34)	(-2.88)	(-2.26)	(-1.97)	(-0.81)	(-0.92)
#Firms	-0.22***	-0.23***	-0.25***	-0.23***	-0.24***	-0.25***
	(-5.97)	(-6.95)	(-7.00)	(-5.78)	(-6.96)	(-6.89)
Data Exposure	1.10***	0.42	0.14	1.14***	0.34	0.12
	(2.80)	(1.48)	(0.53)	(3.64)	(1.04)	(0.41)
Horizon	-16.66***			-16.64***		
	(-33.86)			(-32.16)		
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 generated by StockTwits. The sample includes all available analyst-day-horizon observations between 2005 and 2017. 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. Data Exposure is an aggregate measure of analysts' exposure to data generated by StockTwits, 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. Data Exposure is set to zero prior to StockTwits introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based either on the number of users that have the firm in their watchlist, or the number of hypothetical messages posted 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 Data Exposure 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. 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)					
Data Exposure:		#Watchlist			# Hypothetical Messages	
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Data Exposure × Auto-correlation	1.17***	0.64***	0.58***	0.59***	0.39**	0.35**
	(3.23)	(2.82)	(2.62)	(2.92)	(2.44)	(2.35)
Horizon × Data Exposure	-4.85***	-2.92***	-2.83***	-2.73***	-2.10***	-2.02***
	(-3.88)	(-4.34)	(-4.28)	(-4.18)	(-3.78)	(-3.99)
Horizon × Auto-correlation	0.62*	0.57***	0.55***	0.36*	0.10	0.14
	(1.95)	(3.07)	(3.12)	(1.85)	(0.66)	(0.90)
Data Exposure × Auto-correlation	1.17***	0.64***	0.58***	0.30*	0.31**	0.28**
	(3.23)	(2.82)	(2.62)	(1.77)	(2.31)	(2.09)
Auto-correlation	1.68***	1.74***	1.32***	1.69***	1.75***	1.33***
	(7.38)	(8.88)	(6.66)	(7.21)	(8.73)	(6.57)
Data Exposure	-2.02*	-2.14***	-2.22***	-0.62	-0.99*	-1.13*
	(-1.80)	(-3.25)	(-3.33)	(-0.98)	(-1.66)	(-1.91)
Horizon	-18.07***			-17.97***		
	(-28.26)			(-26.53)		
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 Derivations in the Model

Proof of equation (6). Differentiating $\bar{W}$, given by eq.(3), with respect to f_{st} and f_{lt}, we obtain:

$$\begin{aligned} \frac{\partial \bar{W}}{\partial f_{st}} &= -2\gamma \, \mathsf{E}(f_{st} - \theta_{st} \mid \Omega_1) \\ \frac{\partial \bar{W}}{\partial f_{lt}} &= -2(1-\gamma) \, \mathsf{E}(f_{lt} - \theta_{lt} \mid \Omega_1) \end{aligned} \tag{19}$$

Thus, the first order conditions for the analyst's problem at date 1 yield, using the fact that $\Omega_1 = \{s_{st}, s_{lt}\}$ and that s_{lt} is uninformative about θ_{st}:

$$f^*_{st} = \mathsf{E}(\theta_{st} \mid \Omega_1) = \mathsf{E}(\theta_{st} \mid s_{st}), \tag{20}$$

and

$$f^*_{lt} = \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt}) \tag{21}$$

It is straightforward that the second order conditions for the analyst's optimization problem are satisfied.

Proof of Proposition 1. Substituting $\mathsf{Var}(\theta_{st} \,|s_{st})$ and $\mathsf{Var}(e_{lt} \,|s_{st}, s_{lt})$ in the analyst's objective function in eq.(9) by their expressions in eq.(10), we obtain that the first order conditions for the analyst's optimization problem at date 0 are (ignoring for the moment, the constraints that $\{z_{st}, z_{lt}\} \in [0, Z]^2\}$):

$$\begin{aligned} q(\beta, \gamma)\xi^2_{st} - 2az^*_{st} - cz^*_{lt} &= 0 \\ (1-\gamma)\xi^2_{lt} - 2bz^*_{lt} - cz^*_{st} &= 0 \end{aligned} \tag{22}$$

It is then straightfoward to check that the solution to this system of equations is given by (z^*_{st}, z^*_{lt}) as defined in eq.(11). The Hessian matrix corresponding to the analyst's optimization problem is negative definite and its determinant is positive if and only if $4ab > c^2$. Thus, the solution of the previous system of equations maximizes the analyst's objective function at date 0, provided that $\{z^*_{st}, z^*_{lt}\} \in [0, Z]^2\}$ and $4ab > c^2$.

The condition $z^*_h \leq Z$ is clearly always satisfied by setting Z large enough. Moreover, using the expressions for $\{z^*_{st}, z^*_{lt}\}$ in Proposition 1, it is direct that the condition $z^*_{st} > 0$

is satisfied if and only if:

$$\frac{\xi_{lt}^2}{\xi_{st}^2} \leq \frac{2b \times q(\beta,\gamma)}{c(1-\gamma)},$$

and the condition $z_{lt}^* > 0$ is satisfied if and only if

$$\frac{c \times q(\beta,\gamma)}{2a(1-\gamma)} \leq \frac{\xi_{lt}^2}{\xi_{st}^2}.$$

It is immediate that if these two conditions are satisfied then the condition $4ab > c^2$ is satisfied. Finally, it is easily checked that these two conditions are equivalent to:

$$c < \bar{c}(\beta,\gamma,a,b,\xi_{st}^2,\xi_{lt}^2),$$

where:

$$\bar{c}(\beta,\gamma,a,b,\xi_{st}^2,\xi_{lt}^2) = Min\{\frac{2\frac{\xi_{lt}^2}{\xi_{st}^2}a(1-\gamma)}{q(\beta,\gamma)}, \frac{2bq(\beta,\gamma)}{\frac{\xi_{lt}^2}{\xi_{st}^2}(1-\gamma)}\}.$$

Using the expressions for z_{st}^* and z_{lt}^* in Proposition 1, we deduce that:

$$\begin{aligned} \frac{\partial z_{st}^*}{\partial a} &= -\frac{4b}{(4ab-c^2)}z_{st}^* < 0, \\ \frac{\partial z_{lt}^*}{\partial a} &= \frac{2c}{(4ab-c^2)}z_{st}^* > 0 \quad iff c > 0. \end{aligned} \tag{23}$$

Proof of Corollary 1. Differentiating eq.(13) and eq.(14) with respect to the marginal cost of producing short-term information, a, we obtain

$$\frac{\partial \mathcal{I}_{st}}{\partial a} = (\frac{\partial z_{st}^*}{\partial a})\frac{\xi_{st}^2}{(\kappa_{st}^2 + (Z - z_{st}^*)\xi_{st}^2)^2}, \tag{24}$$

and

$$\frac{\partial \mathcal{I}_{lt}}{\partial a} = (\beta^2\xi_{st}^2\frac{\partial z_{st}^*}{\partial a} + \frac{\partial z_{lt}^*}{\partial a}\xi_{lt}^2)\mathcal{I}_{lt}^2 = -(\frac{2(2\beta^2\xi_{st}^2 b - c\xi_{lt}^2)}{(4ab-c^2)})z_{st}^*\mathcal{I}_{lt}^2. \tag{25}$$

As $\frac{\partial z_{st}^*}{\partial a} < 0$ (see eq.(23)), eq.(24) implies that $\frac{\partial \mathcal{I}_{st}}{\partial a} < 0$. Moreover as $z_{st}^* > 0$, eq.(25) implies that $\frac{\partial \mathcal{I}_{lt}}{\partial a} > 0$ if and only if $\beta < (\frac{c}{2b})^{\frac{1}{2}}\frac{\xi_{lt}}{\xi_{st}}$.

Proof of equations (13) and (14). By definition:

$$\mathsf{Var}(\theta_{lt}\,|f_{lt}^*) = \mathsf{E}((\theta_{lt} - \mathsf{E}(\theta_{lt} \mid f_{lt}^*))^2 \mid f_{lt}^*).$$

As $f_{lt}^* = \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt})$, we deduce that:

$$\mathsf{Var}(\theta_{lt} \,|f_{lt}^*) = \mathsf{E}((\theta_{lt} - \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt}))^2 \mid f_{lt}^*).$$

Using the law of iterated expectations, we deduce that:

$$\mathsf{Var}(\theta_{lt} \,|f_{lt}^*) = \mathsf{E}(\mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})) \mid f_{lt}^*).$$

Now, as $\mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})$ does not depend on the realizations of s_{st} and s_{lt} (due to the assumption that all variables are normally distributed), we obtain that:

$$\mathsf{Var}(\theta_{lt} \,|f_{lt}^*) = \mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt}).$$

Finally, as $\theta_{lt} = \beta\theta_{st} + e_{lt}$, we deduce that:

$$\mathsf{Var}(\theta_{lt} \,|f_{lt}^*) = \beta^2 \,\mathsf{Var}(\theta_{st} \mid s_{st}) + \mathsf{Var}(e_{lt} \mid s_{lt}) + 2Cov(s_{st}, e_{lt} \mid s_{st}, s_{lt})).$$

As s_{lt} and s_{st} are independent and as e_{lt} and θ_{st} are unconditionally independent, we have $Cov(s_{st}, e_{lt} \mid s_{st}, s_{lt}) = 0$. It follows that:

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

and therefore $\mathcal{I}_{lt}$ is as given in eq.(14). The derivation of the expression for $\mathcal{I}_{st}$ follows the same step and is omitted for brevity.

B 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 firms (A, B, C, D, E, F) on December 31, 2006, and forecasting earnings for the fiscal period end 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 firm, 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 firm and the R^2 measure is computed excluding that stock.[46] 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 firm 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 firms (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%.

- 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

[46] 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 firm B

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						
Firm	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

C Variable Definitions

Variable	Definition
All variables below are *analyst-level* variables	
#Firms	Total number of distinct firms covered by an analyst on a given day.
Horizon	Number of days between the date at which the econometrician observes the last available forecasts of the analyst for a given fiscal period, and the date at which actual earnings for each forecast are announced, divided by 365. When earnings announcement date differs across firms covered by the analyst, we use the median date.
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 of this analyst explain a larger fraction of the variation in realized earnings at this horizon.
All variables below are *firm-level* variables that we convert into analyst-level variables by taking the average across all firms the analyst covers	
#Messages	Number of StockTwits' messages posted about a given firm over the last thirty days (from $t-30$ to $t-1$).
#Hypothetical Messages	Number of Hypothetical StockTwits' messages posted about a given firm over the last thirty days (from $t-30$ to $t-1$). The number of hypothetical messages about firm j at time t is computed as $\frac{n_{j,\to t-1}}{N_{\to t-1}} \times N_t$, where $n_{j,\to t-1}$ is the total number of messages posted on StockTwits about firm j since inception until $t-1$, $N_{\to t-1}$ is the total number of messages posted on StockTwits about all firms since inception until $t-1$, and N_t is the total number of messages posted about all firms at time t.
#Watchlist	Total Number of StockTwits' 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 last available financial statements in Compustat).
Cash to assets	che/at (from last available financial statements in Compustat).
Debt to assets	$(dlc+dltt)/at$ (from last available financial statements in Compustat).
Tobin's Q	$(at-ceq+chso*prcc_f)/at$ (from last available financial statements in Compustat).

Table A.2: Data Exposure and News Arrival

The table presents the sensitivity of various measures of social media data to news arrival. This sensitivity is estimated at the firm-day level by OLS. The sample includes all U.S. firms that have been discussed at least once on StockTwits between 2009 and 2017, and that are covered by at least one analyst. In columns (1) to (4), #Watchlist is the number of StockTwits users having the firm in their watchlist on day t. In columns (5) to (8), #Hypothetical Messages is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (9) to (12), #Messages is the number of actual messages posted about the firm from $t-30$ to $t-1$. $\#\text{News}_t$ is the number of distinct news about the firm reported in Capital-IQ - Key Development on day $_t$. $\#\text{News}_{t\to T}$ is the number of distinct news about the firm reported in Capital-IQ - Key Development between day t and day T. Capital IQ - Key Development is a dataset providing structured summaries of material news and events for more than 800,000 firms worldwide. It monitors more than 230 categories of news (i.e., a "key development" item) including for example companies SEC filings, executive changes, M&A announcements, earnings announcements, changes in corporate guidance, delayed filings, SEC inquiries, or credit rating changes. Each "key development item" includes announced date, headline, situation summary, type, company role, and company identifiers. t-statistics in parentheses are based on standard errors clustered by firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	#Watchlist				#Hypothetical Messages				#Messages			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\#\text{News}_t$	-9.86				-2.70				8.32***			
	(-0.94)				(-0.71)				(3.18)			
$\#\text{News}_{t-1}$		-8.81				-1.35				12.67***		
		(-0.84)				(-0.39)				(4.66)		
$\#\text{News}_{t-10\to t-1}$			-7.25				-1.74				13.29***	
			(-0.77)				(-0.54)				(5.43)	
$\#\text{News}_{t-30\to t-1}$				-6.17				-1.35				12.29***
				(-0.73)				(-0.48)				(5.64)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	16'135'311	16'134'268	16'121'531	16'084'251	16'135'311	16'134'268	16'121'531	16'084'251	16'135'311	16'134'268	16'121'531	16'084'251

Table A.3: Social media data and analysts' forecasting activity

This table presents the sensitivity of new forecast issue to StockTwits activity. This sensitivity is estimated by OLS at the analyst-firm-day level. The sample includes all US firms covered by at least one analyst between 2009 and 2017. The dependent variable is a binary variable equal to one if the analyst issues a new forecast (or a revision) on a given firm on day t and zero if not. #Messages is the number of StockTwits' messages posted about a firm from $t-30$ to $t-1$. The number of messages is set to zero when the firm is not covered/discussed on the platform. Trading Volume is the total volume of trading on from $t-30$ to $t-1$. 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 from $t-30$ to t (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 firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Binary Variable (New Forecast=1)			
OLS	(1)	(2)	(3)	(4)
#Messages	0.02***	0.03***	0.06***	0.06***
	(2.97)	(4.29)	(8.82)	(2.70)
Trading Volume		-0.13***	-0.04***	0.09*
		(-9.74)	(-4.12)	(1.86)
Analyst × Firm 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.4: Robustness: Controlling for Trading Activity

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts at different horizons to social media data generated by StockTwits (eq.(18). The sample includes all available analyst-day-horizon observations between 2005 and 2017. 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. Data Exposure is an aggregate measure of analysts' exposure to data generated by StockTwits, 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. Data Exposure is set to zero prior to StockTwits introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the number of users that have the firm in their watchlist, or the number of hypothetical messages posted 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 Data Exposure can be interpreted as the unconditional effect on one-year informativeness. 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)					
Data Exposure:		#Watchlist			# Hypothetical Messages	
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Data Exposure	-1.09***	-1.22***	-1.00***	-0.89***	-1.06***	-0.98***
	(-3.22)	(-3.42)	(-3.74)	(-3.89)	(-4.74)	(-4.89)
Data Exposure	0.16	0.00	-0.30	0.65**	0.33	-0.06
	(0.66)	(-0.02)	(-1.15)	(2.43)	(1.48)	(-0.21)
Horizon × Trading Volume	1.12***	1.19***	0.57***	1.20***	1.28***	0.67***
	(6.56)	(7.58)	(2.67)	(7.09)	(8.18)	(3.42)
Trading Volume	-0.40	-1.19***	-1.23***	-0.50	-1.23***	-1.26***
	(-1.29)	(-3.83)	(-3.80)	(-1.60)	(-4.15)	(-3.98)
Horizon	-17.63***			-17.61***		
	(-31.70)			(-31.31)		
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

Table A.5: 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 generated by StockTwits (eq.(18). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with stable coverage only. Coverage is stable if the level of similarity between the portfolio of firms covered in the current year and that of the previous year is greater than 90%. Similarity is defined as the number of common firms 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 firms 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. Data Exposure is an aggregate measure of analysts' exposure to data generated by StockTwits, 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. Data Exposure is set to zero prior to StockTwits introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the number of users that have the firm in their watchlist, or the number of hypothetical messages posted 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 Data Exposure can be interpreted as the unconditional effect on one-year informativeness. 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)					
Data Exposure:	#Watchlist			#Hypothetical Messages		
OLS	(1)	(2)	(3)	(4)	(5)	(6)
Horizon × Data Exposure	-0.46	-0.64**	-0.69***	-0.17	-0.47***	-0.73***
	(-1.49)	(-1.99)	(-2.60)	(-0.71)	(-2.14)	(-3.49)
Data Exposure	0.32	0.08	-0.15	0.85***	0.40*	0.12
	(1.25)	(0.33)	(-0.52)	(2.95)	(1.66)	(0.44)
Horizon	-16.35***			-16.37***		
	(-36.87)			(-35.33)		
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