Does Alternative Data Improve Financial Forecasting? The Horizon Effect

Olivier Dessaint, Thierry Foucault, and Laurent Frésard*

July 20, 2021

ABSTRACT

We analyze the effect of alternative data on the informativeness of financial forecasts. Our hypothesis is that the emergence of alternative data increases the net benefit of collecting short-term information about firms' cash flows more than the benefit of collecting long-term information. If this hypothesis is correct, and forecasting short-term and long-term cash flows are distinct tasks, alternative data should make short-term forecasts more informative but long-term forecasts less so. We confirm this prediction empirically for sell-side equity analysts' forecasts using a new measure of forecast informativeness at various horizons.

Key words: Alternative data, Equity 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 Tony Cookson, Ahmed Guecioueur, Hazel Hamelin, Gerard Hoberg, Paul Karehnke, Xinyu Liu, Adrien Matray, Randall Morck, Marina Niessner, Gordon Phillips, Eric So, Jerome Taillard, Laura Veldkamp, and participants at the 2021 AEA Meetings, Baruch College, Copenhagen Business School, the Corporate Finance Webinar, Essec, Nova School of Business, McGill University, NBER Big Data and Securities Markets Conference, Neoma Business School, INSEAD, Southern Methodist University, Universita della Svizzera Italiana, University of Amsterdam, University of Laval, University of Geneva, and the 2021 WFA Conference for useful comments. All errors are the authors' alone.

I Introduction

The digitization of information has generated exponential growth in new types of data (e.g., 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 (e.g., financial analysts) forecast future outcomes (e.g., cash flows) and make decisions (e.g., value assets and choose portfolios).[2] However, research on its implications is still 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., investing in stocks or capital budgeting) rely on forecasts that span both short and long horizons.

Alternative data reduces the cost of obtaining information (Goldfarb & Tucker (2019)). In theory, this reduction enables forecasters to obtain more precise signals and form better forecasts (Verrecchia (1982)). Therefore, alternative data should improve the quality of financial forecasting, in general. In this paper, we propose a more nuanced prediction. We conjecture that alternative data predominantly contains information about firms' short-term prospects. For instance, sensor data tracking retailers' activity (e.g., satellite or geolocation data), credit card data, or social media posts about products and brands are informative about firms' upcoming earnings, but less clearly so about earnings in three years' time because firms' long-term prospects depend on their strategic and innovation choices. Anticipating these choices and predicting their implications requires human judgement and more qualitative information (e.g., discussions with industry leaders, scientists, or executives) than that provided by alternative data. Therefore, while alternative data should improve the quality of short-term forecasts, its effect on the quality of long-term forecasts is less clear; it depends on how a decline in the cost of obtaining short-term information affects forecasters' incentives to obtain long-term information.

To shed light on this question, we first consider a model in which a forecaster collects

[1] According to the website Alternative Data.org (http://www.alternativedata.org), there were more than 400 providers of alternative data in 2020 and the amount invested by buy-side investors in such data was close to $2 billion.

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

information to minimize her average forecast errors of the short-term and long-term earnings of a firm. Importantly, we assume that forecasting short-term and long-term earnings entails two distinct tasks. They both require collecting information on the firm's assets in place ("short-term information") but the latter also necessitates information on its growth options ("long-term information"). The forecaster strategically allocates efforts to these tasks and bears a cost for multi-tasking (due, for instance, to cognitive constraints): increasing the effort allocated to one task makes the effort allocated to the other costlier. We show that an increase in the marginal informational return (reduction in forecasting errors) on the effort exerted to obtain short-term information due to a drop in the cost of obtaining short-term information (or an increase in the amount of such information) causes the forecaster to optimally substitute effort away from collecting long-term information. Hence, the availability of alternative data makes her short-term forecasts more informative but it can reduce the informativeness of her long-term forecasts, in particular when the correlation between short-term and long-term earnings is low or the cost of multi-tasking is high.

We test this prediction, focusing on one important type of forecasters, namely sell-side equity analysts. Indeed, to set target prices and make investment recommendations to investors, analysts routinely forecast earnings at short and long horizons, and they do so considering all relevant information, including from alternative data sources.[3] For our tests, we develop a new measure of the informativeness of analysts' forecasts by horizon, exploiting the fact that they typically cover multiple firms. Specifically, we measure the informativeness of the forecasts produced by an analyst on a given day for a given horizon h 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 analyst's forecasts for horizon h explain (in a statistical sense) a larger fraction of the variation in realized earnings at this horizon, and thus that her forecasts are more informative.

Using earnings forecasts from I/B/E/S, we calculate this R^2 every day between 1983 and 2017 for all U.S. analysts and all possible horizons (ranging from one day to five years). We obtain a sample of more than 65 million analyst-day-horizon R^2 observations and find that

[3]See Chi et al. (2021) for evidence that analysts use alternative data and Section 13 in our online appendix for an example.

short-term forecasts are significantly more informative than long-term forecasts. On average, R^2 decreases by 12 percentage points for every one-year increase in horizon. Thus, the term structure of forecasts' informativeness (i.e., the relationship between R^2 and the horizon h) is downward-sloping. Our theory predicts that greater exposure to alternative data increases R^2 for low values of h, but possibly decreases R^2 for high values of h, thereby rendering the slope of the term structure of analysts' forecast informativeness steeper.

Testing this prediction requires identifying variation in analysts' exposure to alternative data, which is empirically challenging for three reasons. First, "alternative data" is a generic term for any data containing relevant information about firm value that is not directly disclosed by firms, and refers to a myriad of datasets.[4] Second, these datasets greatly vary in their scope and most of them are only relevant for a subset of firms (e.g., credit card data are informative for retail activities but less so for steel manufacturing). Last, variation in analysts' exposure to alternative data may be related to confounding factors also affecting their R^2. Thus, the key empirical challenge is to identify a source of variation in alternative data that is (i) common to hundreds of alternative data providers, (ii) relevant for the entire cross-section of firms, and (iii) unrelated to other factors affecting R^2. Since building a single test satisfying all three conditions is difficult, we develop two tests, one at the "macro-level" that meets the first two conditions, and another one at the "micro-level" that plausibly meets the last two.

[Insert Figure I about here]

Our first, "macro-level", test exploits the rise in the number of alternative data providers over time, particularly since the late 2000s (see Figure I). We ask whether this trend (common to all alternative data and relevant for all firms) coincides with an increase in the informativeness of short-term analysts' forecasts and a decrease in the informativeness of their long-term forecasts, as predicted by our model. We find that those two opposite effects are indeed present in the data. On average, R^2 at the one-year horizon has increased by roughly 10 percentage points since 2000 from about 60% to 70%, but decreased at the five-year horizon by roughly 10 percentage points from about 40% to 30%. We also show

[4]See Section 12 in the online appendix for a taxonomy of available alternative data as of this writing.

that the annual "slope" of the term structure has become steeper over time, a trend that has accelerated since 2005. Since other factors may also explain this long-run evolution of the term structure, this first test does not provide causal evidence in support of our theory. Nevertheless, it provides estimates that serve as a benchmark for gauging the aggregate effect of the rise of alternative data.

Our second, "micro-level", test focuses on social media data. We study the effects of analysts' exposure to data generated by StockTwits, a social networking platform where investors share information (blog posts, charts, or links to articles about a stock) and opinions about individual firms.[5] StockTwits is well-suited for our purpose because the data generated therein is (i) social media data – an important source of alternative data – covering almost every firm (in contrast to many other alternative datasets, whose coverage is more specialized), (ii) contains (as we show) information mostly relevant about firms' short-term cash flows, and (iii) is used by analysts (we provide evidence thereof). Moreover, StockTwits was introduced in 2009 and expanded progressively with different level of intensity across firms. This feature enables us to estimate the effect of greater exposure to alternative (social media) data using an approach similar to a difference-in-differences, namely by comparing how R^2 (for a given horizon) changes after the introduction of the StockTwits platform for analysts with early and high exposure to StockTwits data relative to analysts who were exposed later or simultaneously but with less intensity.

We measure analysts' exposure using two complementary approaches, aimed at capturing variations in data generated on StockTwits that would not have been available to analysts from traditional sources (e.g., company filings or press releases). Our first measure is the daily average number of users who have on their "watchlist" (i.e., the list of firms they follow) the firms covered by the analyst. Since users rarely modify their watchlist after registering on StockTwits, a firm's watchlist (i.e., the number of users having the firm in their watchlist) changes because new users register and enter the platform. Therefore, variation in the number of users on a firm's watchlist mostly reflects the overall expansion of StockTwits,

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

both over time and across firms, and not the arrival of information from other sources. Our second measure 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) correspond to the total number of messages on StockTwits (across all firms) multiplied by the focal firm's average share of total messages.[6] As this share is constant, the number of hypothetical messages about a firm (unlike actual messages) does not change with the arrival of firm-specific information from other sources (which we also confirm empirically). Both measures are set to zero before 2009 and are used as the main explanatory variable in a specification controlling for analyst and time fixed effects, which we estimate separately by horizon sub-sample over the 2005-2017 period.

For both measures, we find that greater exposure to alternative (social media) data is associated with a significant improvement in the informativeness of analysts' short-term forecasts (less than one year), and a decline of comparable magnitude in the informativeness of their long-term forecasts (beyond two years). We also find that the slope of the term structure of forecast informativeness becomes steeper for more exposed analysts, and even more so for exposed analysts whose name matches that of a StockTwits user account. This steepening is also more pronounced for (i) analysts following more firms (i.e., those for whom the cost of multi-tasking is plausibly higher), and (ii) analysts following firms whose earnings are less autocorrelated. Therefore, the steepening of the term structure of forecast informativeness due to increased exposure to alternative data varies systematically across analysts, as our theory predicts.

The rest of the paper is organized as follows. Section II positions our contribution in the related literature. In Section III, we present our model of forecasting by analysts and derive our main prediction. Section IV presents the data used in our tests and our new measure of analysts' forecast informativeness. In Sections V and VI, we report the findings of our macro- and micro-level tests of the effects of alternative data on the informativeness of analysts' forecasts. Section VII concludes. All derivations for our model and the definitions

[6]Intuitively, "hypothetical messages" is 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 an average day.

for the variables used in our tests are reported in the Appendix.

II Contribution to the Literature

Our results add to a growing literature 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 about firms' fundamentals (or relaxes information capacity constraints) and study the implications for the informativeness of asset prices (Dugast & Foucault (2018), Farboodi & Veldkamp (2020)), market efficiency (Martin & Nagel (2020)), firms' growth rates (Begeneau et al. (2018)), information acquisition choices by asset managers (Abis (2018)), the pricing of information by data vendors (Huang et al. (2020)), or financial inclusion (Mihet (2020)).

We also assume that the emergence of alternative data reduces the cost of information acquisition. However, in contrast to the literature, we consider the possibility that this reduction is *heterogeneous* across horizons. We conjecture that most types of alternative data contain short-horizon information about fundamentals and therefore reduce the cost of obtaining information about short-term cash flows significantly more than about long-term cash flows. To our knowledge, our paper is the first to formulate this hypothesis and analyze its implications for the informativeness of financial forecasts when forecasters face a trade-off between collecting short-term and long-term information. This trade-off is relevant because most financial decisions require making forecasts about outcomes that occur at different dates in the future.[7]

Our focus on the informativeness of financial forecasts differentiates our study from existing papers analyzing the effects of digitization and alternative data on the informativeness of order flows and asset prices. Using various sources of variation – the digitization of firms' regulatory filings (Gao & Huang (2020)), the availability of satellite images of retailers' parking lots (Zhu (2019)) or variations in the volume of data generated by financial blog posts (Grennan & Michaely (2020*b*)) – these papers conclude that digitization or alternative data

[7]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).

make order flow (e.g., the net order imbalance from individual investors or the absolute order imbalance) and stock prices more informative (e.g., the predictability of future earnings announcements based on stock returns or measures of price non-synchronicity).[8]

This conclusion does not contradict our finding that alternative data negatively affects the informativeness of analysts' long-term forecasts. Indeed, as stock prices are the sum of discounted forecasted cash flows at all horizons, their informativeness about firms' cash flows at specific future dates depends on the informativeness of both short-term and long-term forecasts. Thus, the net effect on the informativeness of prices of an improvement in the informativeness of short-term forecasts and a deterioration in that of long-term forecasts due to alternative data is ambiguous. This ambiguity is difficult to address given the challenge of measuring the horizons at which stock prices provide useful information. Gao & Huang (2020), Grennan & Michaely (2020*b*) and Zhu (2019) provide evidence consistent with digitization or alternative data making prices more informative about short-horizon earnings (i.e., up to one year), but do not consider longer horizons. To the extent that analysts' forecasts are representative of those of investors, their results are consistent with our findings that alternative data enhances the informativeness of short-term forecasts (less than 1 or 2 years).

Relatedly, Bai et al. (2016) and Farboodi et al. (2020) ask whether long run reductions in information processing costs due to progress in information technology and data abundance have changed the informativeness of stock prices. Interestingly, they report ambiguous effects, with improvements for some firms (e.g., large firms) but deteriorations for others (e.g., small firms). Our findings suggest that this heterogeneity could be related to the evolution of the informativeness of analysts' forecasts at various horizons (a question which is beyond the scope of our paper). In any case, given the importance of earnings forecasts for asset valuation and capital allocation, these results call for a detailed analysis of the effects of alternative data on the quality of financial forecasting at short and long horizons. Our paper is a first step in this direction.

[8]These findings suggest that alternative data contains information. Several papers establish that this is the case by assessing whether different types of alternative data help in predicting stock returns and firms' fundamentals (e.g., sales, and earnings). For instance, social media posts (Chen et al. (2014)), product reviews (Huang (2018) and Tang (2018)), employees' reviews (Green et al. (2019)), online customers' activity (Froot et al. (2017)) or satellite images (Katona et al. (2021), Mukherjee et al. (2021)) have been found to contain information about stock returns and fundamentals.

Finally, our paper also contributes to the literature studying how analysts form their forecasts. To our knowledge, our finding of a downward sloping term structure of equity analysts' forecasts' informativeness and its evolution over time is novel. In addition, our findings add to existing research studying the determinants of analysts' effort allocation (Harford et al. (2019) or Hirshleifer et al. (2019)), the properties and implications of short-term and long-term forecasts (Bandyopadhyay et al. (1995) or Mest & Plummer (1999)), and how progress in information technologies affects the organization and output of the financial analysis industry (Gerken & Painter (2021), Chi et al. (2021), van Binsbergen et al. (2020), or Grennan & Michaely (2020*a*)).

III Hypothesis Development

A The Forecasting Problem

The model features one forecaster (the "analyst") and one firm. Figure II presents the timeline. The firm generates two cash flows (earnings), θ_{st} at date 2 (the short-term) and θ_{lt} at date 3 (the long-term) with

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

where $\beta \geq 0$, $\theta_{st} \sim \mathcal{N}(0, \sigma_{st}^2)$, $e_{lt} \sim \mathcal{N}(0, \sigma_e^2)$, and θ_{st} and e_{lt} are independent. The long-term earnings are the sum of two components: (i) the *common component* ($\beta\theta_{st}$) generated, for instance, by assets in place, and (ii) *the unique component* (e_{lt}) generated, for instance, by growth opportunities. The correlation between the long-term and short-term earnings increases with β.

[Insert Figure II about here]

At date 1, the analyst formulates forecasts about θ_{st} and θ_{lt} denoted f_{st} (the short-term forecast) and f_{lt} (the long-term forecast), respectively. Her payoff, denoted $W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st})$, is inversely related to the weighted sum of her squared forecast 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 $0 < \gamma < 1$ (see Section III.C for a discussion).

To generate her forecasts, the analyst uses a "short-term signal", s_{st}, about the short-term earnings and a "long-term signal", s_{lt}, about the unique component (e_{lt}) of the long-term earnings:

$$s_{st} = \theta_{st} + \epsilon_{st}, \quad s_{lt} = e_{lt} + \epsilon_{lt}, \tag{3}$$

where $\epsilon_h \sim \mathcal{N}(0, Z_h - \psi_h z_h)$ and $z_h \leq (\psi_h)^{-1} Z_h$, for horizon $h \in \{st, lt\}$. The errors (ϵ_hs') in the analyst's signals are independent from each other and all other variables in the model. Variable z_h denotes the analyst's effort (chosen at date 0; see below) to collect and process information at horizon h. An increase of z_{st} (resp., z_{lt}) reduces the variance of the noise in her short-term (long-term) signal, s_{st} (resp., s_{lt}), at rate ψ_{st} (resp., ψ_{lt}). Thus, ψ_h measures the "informational return" of effort at horizon h. The effort exerted to improve the precision of one signal does not affect the precision of the other one because the efforts required to collect information about the common and unique components of future earnings are distinct tasks (the unique component is not correlated with the common component). For given forecasts $\{f_{st}, f_{lt}\}$, the analyst's expected payoff conditional on her information at date 1 is

$$\begin{aligned} \overline{W}(f_{st}, f_{lt}; s_{st}, s_{lt}) &\equiv \mathsf{E}(W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st}) \,|s_{st}, s_{lt}) \\ &= \omega - \gamma\, \mathsf{E}((f_{st} - \theta_{st})^2 \,|s_{st}, s_{lt}) - (1-\gamma)\, \mathsf{E}((f_{lt} - \theta_{lt})^2 \,|s_{st}, s_{lt})\,. \end{aligned} \tag{4}$$

The analyst chooses her optimal forecasts, $\{f^*_{st}, f^*_{lt}\}$, to maximize $\overline{W}(f_{st}, f_{lt}; s_{st}, s_{lt})$. Thus (see Appendix III for all derivations)

$$f^*_{st} = \mathsf{E}(\theta_{st} \,|s_{st})\,, \quad f^*_{lt} = \mathsf{E}(\theta_{lt} \,|s_{st}, s_{lt})\,. \tag{5}$$

Substituting eq.(5) into eq.(4), we obtain that the analyst's unconditional (date 0) expected payoff is

$$\mathsf{E}(\overline{W}(f^*_{st}, f^*_{lt}; s_{st}, s_{lt})) = \omega - q(\beta, \gamma)\, \mathsf{Var}(\theta_{st} \,|s_{st}) - (1-\gamma)\, \mathsf{Var}(e_{lt} \,|s_{lt})\,, \tag{6}$$

where $q(\beta, \gamma) \equiv \gamma + (1-\gamma)\beta^2$. The analyst's *expected* payoff increases in the informativeness of her signals (the inverse of $\mathsf{Var}(\theta_{st} \,|s_{st})$ and $\mathsf{Var}(e_{lt} \,|s_{lt})$) because more informative signals reduce her average forecast errors. The informativeness of the signal at horizon h increases in the analyst's effort to obtain information specific to this horizon because, as explained

previously, this effort reduces the noise in the analyst's signal. For instance, if the analyst's priors about θ_{st} and e_{st} are diffuse then

$$\mathsf{Var}(\theta_{st}\,|s_{st}) = (Z_{st} - \psi_{st} z_{st}), \quad \mathsf{Var}(e_{lt}\,|s_{lt}) = (Z_{lt} - \psi_{lt} z_{lt}). \tag{7}$$

Exerting effort is costly for the analyst. Her total information processing cost is

$$C(z_{st}, z_{lt}) = C_0 + a \times z_{st}^2 + b \times z_{lt}^2 + c \times z_{st} z_{lt}, \tag{8}$$

where C_0 is the fixed cost of understanding the firm's business and collecting information about it. As usual in the literature on information acquisition, 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. Furthermore, we assume that multi-tasking is costly, $c > 0$. For instance, if the analyst already exerted a lot of effort to improve the precision of, say, her short-term signal then it becomes more taxing to exert even more effort, be it to improve the precision of the short-term signal ($a > 0$) or the precision of the other signal ($c > 0$).[9]

The analyst chooses her efforts, z_{st}^* and z_{lt}^*, at date 0 to maximize her ex-ante expected payoff net of the cost of effort, $J(z_{st}, z_{lt}) \equiv \mathsf{E}(\overline{W}(f_{st}^*, f_{lt}^*; s_{st}, s_{lt})) - C(z_{st}, z_{lt})$. Thus, z_{st}^* and z_{lt}^* solve

$$\max_{z_{st} \leq (\psi_{st})^{-1} Z_{st}, z_{lt} \leq (\psi_{lt})^{-1} Z_{lt}} J(z_{st}, z_{lt}) = \omega - q(\beta, \gamma)\, \mathsf{Var}(\theta_{st}\,|s_{st}) - (1-\gamma)\, \mathsf{Var}(e_{lt}\,|s_{lt}) - C(z_{st}, z_{lt}). \tag{9}$$

In choosing her efforts, the analyst trades off the precision of her signals against the cost of effort. To solve for her optimal efforts, it is analytically convenient to assume that the analyst's priors about θ_{st} and e_{st} are diffuse ($\mathsf{Var}(\theta_{st}\,|s_{st})$ and $\mathsf{Var}(e_{lt}\,|s_{lt})$ are given by eq.(7)). In this case, we obtain the following result.[10]

Proposition 1 *: When $c \leq \bar{c}(\beta, \gamma, a, b, \psi_{st}, \psi_{lt})$ (where $\bar{c}$ is defined in the proof of the proposition) and Z_h is large enough for $h \in \{st, lt\}$, the analyst's optimal efforts in producing*

[9] Goldstein & Yang (2015) consider a model in which the payoff of an asset is the sum of three components 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 $c > 0$.

[10] We assume that ω is large enough so that it is always optimal for the analyst to pay the fixed cost C_0 of coverage (i.e., $J(0,0) > 0$).

information at date 0, z_{st}^* *and* z_{lt}^*, *are interior (i.e.,* $0 < z_h^* < (\psi_h)^{-1} Z_h$*) and given by*

$$z_{st}^* = \frac{2bq(\beta,\gamma)\psi_{st} - c(1-\gamma)\psi_{lt}}{4ab - c^2} \quad z_{lt}^* = \frac{2a(1-\gamma)\psi_{lt} - cq(\beta,\gamma)\psi_{st}}{4ab - c^2}. \tag{10}$$

When the marginal cost of producing the short-term signal, a*, decreases then the analyst increases her effort (*z_{st}^**) to improve the precision of her short-term signal and, if* $c > 0$*, decreases her effort (*z_{lt}^**) to improve the precision of her long-term signal.*

A reduction in the marginal cost of obtaining short-term information raises the net marginal benefit of improving the precision of her short-term signal for the analyst. As a result, the analyst reacts by exerting more effort to obtain short-term information. This reaction is optimal but it raises the marginal cost of exerting effort to improve the precision of the long-term signal because multi-tasking is costly ($c > 0$). Consequently, the analyst optimally reduces the effort she allocates to this task. This mechanism can work either via a reduction in the cost of obtaining short-term information (as here) or an increase in the informational return on effort to obtain short-term information (ψ_{st}), because what matters is the change in the marginal benefit of efforts allocated to each task (see Section III.C).

B Alternative Data and Forecasts' Informativeness

Our hypothesis is that alternative data mainly contains short-term information. Hence, it reduces the marginal cost of obtaining short-term information (or the informational return on effort to obtain short-term information). If this hypothesis is correct, Proposition 1 implies that the emergence of alternative data should lead forecasters using such data to exert (even) more effort to improve the precision of their short-term signals at the expense of the precision of their long-term signals. Forecasters' efforts are not directly observable. However, as shown in Corollary 1, one can use the informativeness of their forecasts to test the implications of Proposition 1.

Intuitively, the analyst's forecast at horizon h is more informative if the residual uncertainty about earnings at this horizon after observing the analyst's forecast, $\mathsf{Var}(\theta_h \mid f_h^*)$, is smaller. Hence, we define the informativeness of the analyst's forecast at horizon $h \in \{st, lt\}$,

denoted by $\mathcal{I}_h$, as the inverse of $\mathsf{Var}(\theta_h \mid f_h^*)$:

$$\mathcal{I}_h \equiv \mathsf{Var}(\theta_h \mid f_h^*)^{-1} \quad \text{for} \quad h \in \{st, lt\}. \tag{11}$$

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 III):

$$\mathcal{I}_{st} = \mathsf{Var}(\theta_j \mid s_{st})^{-1} = (Z_{st} - \psi_{st} z_{st}^*)^{-1}. \tag{12}$$

and

$$\mathcal{I}_{lt} = \mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})^{-1} = (\beta^2 (Z_{st} - \psi_{st} z_{st}^*) + (Z_{lt} - \psi_{lt} z_{lt}^*))^{-1}. \tag{13}$$

The informativeness of the analyst's short-term forecast depends only on her optimal effort (z_{st}^*) to collect information about the common component of the firm's future earnings and 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}^*) because information about the common component is also useful to forecast the long-term earnings (as long as $\beta > 0$).

Corollary 1 *: If $\beta < (\frac{c\psi_{lt}}{2b\psi_{st}})^{\frac{1}{2}}$, 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.*

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 precision of the short-term signal and less effort into increasing the precision of the long-term signal. The first effect raises the informativeness of the long-term forecast while the second reduces it. Corollary 1 shows that the second effect dominates when the correlation between the long-term and short-term earnings is low or when the cost of multi-tasking is large enough ($\beta < (\frac{c\psi_{lt}}{2b\psi_{st}})^{\frac{1}{2}}$).[11] 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.

[11]The condition $\beta < (\frac{c\psi_{lt}}{2b\psi_{st}})^{\frac{1}{2}}$ 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.

These differential effects lead to our main testable implication. Insofar as alternative data increases the marginal net benefit of effort exerted for obtaining short-term information (e.g., via a decrease in the marginal cost of obtaining short-term information), its availability should be associated with an increase in the informativeness of short-term forecasts and a *decrease* in the informativeness of long-term forecasts, especially for firms with a relatively low autocorrelation of earnings (low β) or analysts for which the cost of multi-tasking (c) is high.

C Discussion and Interpretation

Analysts' Objective Function. We test the previous implications using sell-side equity analysts' forecasts. As assumed in our model (see eq.(2)), equity analysts care about the accuracy of their forecasts because their career outcomes (compensation and upward mobility) are positively related to this accuracy (i.e., inversely related to their forecast errors). 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. This relationship might be direct (i.e., the analyst's compensation explicitly depends on her forecast errors) or indirect, when the analyst's career depends on the quality of her recommendations or the validity of the price target that she sets for a stock based on her forecasts.

In reality, analysts issue long-term forecasts less frequently than short-term forecasts (see Section IV.A). However, this fact does *not* imply that they only care about the accuracy of their short-term forecasts (the case $\gamma = 1$ in our model). Indeed, to make investment recommendations or set price targets, analysts must forecast earnings at different future dates. The quality of their investment recommendations is therefore dependent on the accuracy of their short and long-term forecasts. In fact, the literature shows that analysts' long-term forecasts have the greatest explanatory power for analysts' recommendations (Bradshaw (2004)), and that the market reaction to these recommendations is stronger when they are accompanied by long-term forecasts (Jung et al. (2012)). Moreover, revisions in long-term forecasts induce strong market reactions (Chen et al. (2013), Da & Warachka (2011), or Copeland et al. (2004)), suggesting that those forecasts matter for investors, which supports our assumption that the accuracy of long-term forecasts is relevant for analysts' careers. (i.e., $\gamma < 1$ in

(eq.(2))).

Splitting Tasks. When the cost of obtaining short-term information drops, the analyst in our model reallocates effort to the task of collecting short-term information. This behavior is optimal for the analyst (i.e., it minimizes her average total forecasting error) because it saves on the cost of multitasking but it can reduce the accuracy of her long-term forecasts. One may then wonder whether the analyst (or her employer) could not be better off by dividing the tasks of forecasting short-term and long-term earnings between two forecasters. We show in Section 1 of the online appendix that this not the case when $c \leq \frac{4C_0}{q(\beta,\gamma)(1-\gamma)\psi_{st}\psi_{lt}}$. Indeed, under this condition, the increase in fixed costs of information production (each forecaster bears the fixed cost C_0 of collecting information to understand the firm's business) cancels out savings on the cost of multi-tasking. Other frictions (agency and communication costs; see the online appendix) can also explain why splitting the tasks of forecasting short and long-term earnings between two agents is not optimal, even when $C_0 = 0$.

Alternative interpretation. Instead of reducing the cost of obtaining short-term information, alternative data can increase the informational return on effort for obtaining short-term information, i.e., ψ_{st} (e.g., because one can use quantitative tools to analyze alternative data). Our main prediction in this case is unchanged (see Section 2 in the online appendix). In particular, if $\beta < \frac{1}{2}(\frac{c\psi_{lt}}{b\psi_{st}})^{\frac{1}{2}}$ (a condition qualitatively similar to that in Corollary 1), an increase in ψ_{st} improves the informativeness of the analyst's short-term forecast and reduces the informativeness of her long-term forecast. What matters for our prediction is therefore that the emergence of alternative data raises the marginal benefit of effort exerted for obtaining short-term information and not the exact channel for this effect.

IV Measuring Forecasts' Informativeness

A Earnings Forecasts and Realizations

We build a large sample of analyst forecasts of earnings per share (EPS) and net income (in US dollars) from the I/B/E/S Detail History File (Adjusted and Unadjusted) at different horizons (up to 5 years). We exclude quarterly and semi-annual earnings forecasts, and

retain annual earnings forecasts associated with a well-defined fiscal period.[12] We eliminate forecasts with missing announcement dates, analyst code, or broker code. When an analyst issues multiple forecasts for a firm and horizon on a given day, we keep the last forecast based on the 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.[13] 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 only the actual EPS is reported, we convert it into actual net income using the fully diluted number of shares from Compustat if the firm does not have multiple share classes or the number of shares from CRSP if not. Last, we require that (i) actual earnings and total assets at the end of the forecasted fiscal period are not missing, and the absolute value of the former is not greater than the latter, (ii) all forecasts are about a fiscal year ending between 1983 to 2017, (iii) the forecast is issued before the actual earnings report date and this report date occurs after the end of the forecasted fiscal period, and (iv) forecasts (in absolute value) are not greater than 10 times total assets at the end of the forecasted fiscal period. We obtain 9,129,282 unique forecasts and realizations by analyst-firm-date-horizon, which we use next to build our measure of forecast informativeness.

This sample contains 4,259,465 million forecasts with horizon less than one year and 1,260,796 million with horizon greater than two (including 102,431 beyond 4 years), where horizon is the number of days between the forecast date and the earnings report date, divided by 365. Three factors explain why there are more short-term than long-term unique forecasts. First, for inclusion in our sample, the earnings realization must be non-missing. As the

[12] 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.

[13] If an analyst simultaneously issues a net income and EPS forecast, 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 immediate because I/B/E/S does not report the number of shares used by the analyst to make the EPS forecast. Based on instances where we observe both an EPS and a net income forecast, we find that the approach minimizing 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 (see Section 3 in the online appendix).

horizon increases, some firms become inactive before we observe the corresponding earnings. Second, for horizons beyond two years, many analysts only disclose a forecast about long-term growth without explicit horizon.[14] In those cases, the horizon is missing, although the analyst did express a view about the long-term. Last, updating (and consequently reporting) frequency decreases with horizon. Short-term forecasts – which are for the current fiscal year – are regularly updated (or reiterated) before and after quarterly reports, whereas updates of long-term forecasts mostly occur after annual earnings announcements.

As discussed in Section III.C, the imbalance between the number of long and short-term forecasts in I/B/E/S does *not* imply that the trade-off highlighted in our theory is irrelevant and that analysts do not care about long-term forecasts. Arguably, analysts and firms for which we observe long-term forecasts might be different. To mitigate concerns about self-selection, we will verify that our main results hold for analysts disclosing both short and long-term forecasts, and for firms with both types of forecasts.

B Measuring Forecast Informativeness with R^2

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

$$e_j = k_0 + k_1 \hat{e}_j + \nu_j, \tag{14}$$

where j indexes all firms covered by analyst i at time t with an 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. By definition, the R^2 of eq.(14) is

$$R^2_{i,t,h} = 1 - \frac{\mathsf{Var}(\nu_j)}{\mathsf{Var}(e_j)} = 1 - \frac{\mathsf{Var}(e_j|\widehat{e_j})}{\mathsf{Var}(e_j)}. \tag{15}$$

A higher $R^2_{i,t,h}$ implies that analyst i's forecasts at horizon h on day t explain a larger fraction of the variation in realized earnings at date $t+h$ across the firms she covers. Thus, a higher $R^2_{i,t,h}$ indicates that analyst i's forecasts at horizon h are more informative.

Appendix I describes the detailed procedure to compute R^2 for an analyst on a given day and horizon. We apply this procedure to all analysts at all dates between January 1, 1983

[14] See Section 4 in the online appendix.

and December 31, 2017, and for all possible horizons between 1 day to 5 years. Our final sample contains 65,889,122 analyst-day-horizon observations of $R^2_{i,t,h}$, obtained from 14,379 distinct analysts who issued forecasts about 13,849 distinct firms.

C Why R^2 to Measure Forecast Informativeness?

Our measure of analyst forecast informativeness, R^2, follows from our theoretical measure of informativeness, $\mathcal{I}$ (see eq.(11)). Indeed, in the model, the theoretical R^2 of a regression of the earnings at horizon h, θ_h, on the analyst's forecast f_h is $R^2_h = 1 - \mathsf{Var}(\theta_h \mid f_h)/\mathsf{Var}(\theta_h)$. Thus, when the informativeness of the analyst's forecast ($\mathcal{I}_h$) is higher in theory (i.e., when $\mathsf{Var}(\theta_h \mid f_h)$ is smaller), R^2_h is higher.[15] R^2 relies on the intuition that a signal for a given horizon (f_h) is more informative if observing it reduces the residual uncertainty about the uncertain outcome (θ_h) by a larger amount *relative* to prior uncertainty, i.e., if R^2_h is higher. Note that R^2_h is a normalized measure (so that it is comparable across analysts) scaled between 0 (the analyst's forecast is only noise) and 1 (the analyst has perfect foresight).

R^2 is similar in spirit to the measure of price informativeness developed by Bai et al. (2016), but it is new in the literature on analysts, which typically measures informativeness using either analysts' absolute (or squared) forecast error (often called "accuracy"), or the impact of analysts' forecasts on stock prices (Hilary & Hsu (2013) or Merkley et al. (2017)). Using R^2_h rather than these measures has several advantages for our purpose. First, it accounts for the intrinsic difficulty of forecasting by normalizing the residual uncertainty about earnings at horizon h after observing an analyst's forecasts at this horizon ($\mathsf{Var}(\theta_h \mid f_h)$) by a measure of prior uncertainty ($\mathsf{Var}(\theta_h)$). This is important because the residual uncertainty can vary across analysts and within analysts over time, either because of variations in analysts' efforts to collect information or variations in uncertainty about earnings (e.g., uncertainty is higher during recessions; see Bloom (2014)). We are interested in the first source

[15] R^2_h is analyst-specific while in the model it is analyst- and firm-specific. In estimating eq.(15), we are treating each pair $(e_j, \hat{e}_j)$ for fixed values of h, t, and i as different realizations of the pair (θ_h, f^*_h) in the model. Our assumption is that firms' normalized earnings (e_j) and forecasts ($\hat{e}_j$) in a given analyst's portfolio are realizations of the same underlying (gaussian) distributions, and thus that analysts cover firms with similar characteristics. Building a measure that is analyst- *and* firm-specific as in the model, i.e., $R^2_{i,j,t,h}$ where j would index firm is impossible, because this requires multiple forecasts (and realizations) for the same firm by the same analyst for the same horizon at the same time.

of variation, not the second one. In contrast to R^2, absolute (or squared) forecast errors at a given horizon can be large because analysts exert little effort to collect information relevant for this horizon or because uncertainty at this horizon is large (or both).

Second, the absolute (or squared) error can also be large (and yet the forecast still be informative) when analysts are systematically biased, maybe due to conflicts of interest (Hong & Kacperczyk (2010)). In contrast, R^2 is not affected by the average level of the analyst's bias and is identical to the informativeness of analysts' debiased forecasts if the analyst's bias is constant across the firms she covers at time t (see Section 5 in the online appendix). Thus, R^2 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 and short-term forecasts at the same time. This coincidence precludes building a market-based measure of forecast informativeness for a specific *horizon* because one cannot disentangle the contribution of each forecast to the price reaction.

D R^2: Summary Statistics and Stylized Facts

Table I presents summary statistics for R^2. For the 1983-2017 period, an analyst's earnings forecasts explain 68% of the variation in realized earnings across the firms she covers (the average R^2 is 68.01%). The average horizon of her forecasts is 1.11 years, and she typically covers 8.12 firms.

[Insert Table I about here]

The average R^2 decreases with horizon.[16] It is 79.60% for horizons shorter than one year, 59.21% for horizons between one and two years, 49.37% between two and three, 37.62% between three and four, and 31.18% beyond four years. We refer to the relationship between R^2 and h as the term structure of forecasts' informativeness, or simply the "term structure".

[Insert Figure III about here]

[16] We have fewer observations of R^2 at long horizons because we have fewer forecasts at long horizons, and mechanically so at the end of the sample.

To better characterize the shape of this term structure, we plot the means of analyst-level $R^2_{i,t,h}$ over all i and t, by horizon h in number of months (displayed on the x-axis). Figure III (Panel A) confirms that the term structure is downward-sloping. The slope of this term structure, estimated by regressing the means of R^2 by h on h, is negative and equal to -1 (t-stat=-24). Its intercept is 81 (t-stat=54). This linear approximation implies that informativeness (R^2) deteriorates by 1 percentage point for every one-month increase in horizon, i.e, 12 percentage points per year.

E Testing the Effect of Alternative Data on R^2

Our theory implies that greater exposure to alternative data increases $R^2_{i,t,h}$ for low values of h, but possibly decreases $R^2_{i,t,h}$ for high values of h, thereby steepening the *slope* of the term structure. To test this theory, we thus need to characterize the evolution of $R^2_{i,t,h}$ at various horizons. We use two approaches for that. In the first approach, we focus on the *level* of R^2 by horizon and study changes in $R^2_{i,t,h}$ separately for fixed values of h. In the second approach, we focus on the *slope* of the term structure, which we estimate by linear approximation.

Next, for both approaches, we need to identify variation in analysts' exposure to alternative data. In doing so, we face three challenges. First, the term "alternative data" is generic and refers to any data containing relevant information about firms' fundamentals that is not directly disclosed by firms (see Section 12 in the online appendix for a taxonomy of alternative data). A myriad of datasets, introduced progressively over the last thirty years, corresponds to this definition. Second, these datasets greatly vary in their scope and most of them are only relevant for a subset of firms (e.g., credit card data are informative for retail but less so for steel manufacturing). Last, variation in analysts' exposure to alternative data may be related to confounding factors also affecting R^2. Hence, testing the effect of "alternative data" on the informativeness of analysts' forecasts requires finding a source of variation that is (i) common to all alternative data, (ii) relevant for a large set of firms covered by analysts, and (iii) unrelated to other factors affecting R^2. Building a single test that jointly satisfies all three conditions is difficult. For this reason, we combine two complementary tests: a "macro-level" test that satisfies the first two conditions, and a "micro-level" test

that plausibly satisfies the last two.

The first test (Test#1) focuses on the aggregate evolution of the term structure of forecasts' informativeness. Indeed, a common feature of all alternative data sources is that (almost) none of them existed 30 years ago. The first alternative data emerged in the early 90's (see Figure I). Their number and variety then expanded every year, together with their coverage of the cross-section of firms. We thus posit that analysts have become progressively more exposed to all sources of alternative data. Based on our model, this increased exposure should lead to a steepening of the term structure over time. Of course other factors may also affect the evolution of the term structure, so this test cannot provide causal evidence in support of our hypothesis. However, it provides a benchmark for gauging the aggregate effect of the rise of alternative data on financial forecasting.

The second test (Test#2) focuses on analysts' specific exposure to one major source of alternative data: social media data. We use data generated by StockTwits, a social network of traders posting messages about individual stocks. We exploit the heterogeneous expansion of this platform across stocks to construct two measures of analysts' exposure to social media data capturing variation in one important source of alternative data that is relevant for the entire cross-section of firms and plausibly unrelated to other forces affecting R^2.

V Test#1: Long-Run Evolution

This section investigates whether the term structure has become steeper over time, as predicted by our theory.

A Forecast Informativeness by Horizon

To first illustrate the evolution of the term structure of forecasts' informativeness, we split our sample into two sub-samples covering periods of equal length (1983-1999 and 2000-2017) and compare the average term structure over each period. Panel B of Figure III shows that it is steeper in the second half of the sample. This steepening is consistent with our main prediction but it could be due to a structural change around the year 2000.[17] To verify

[17]For example, Srinidhi et al. (2009) document an improvement in the precision of the idiosyncratic information component in short-term forecasts in the two years following regulation Fair Disclosure (FD) in

that this steepening corresponds to a general trend, we compute and plot the means of $R^2_{i,t,h}$ by year, separately for short ($h < 1$) and long-term ($h \geq 2$) forecasts. Figure IV visually confirms the presence of two opposing trends: an improvement in R^2 for short-term forecasts, and a deterioration for long-term ones.

[Insert Figure IV and Table II about here]

To formally test whether these opposite trends are statistically significant, we regress $R^2_{i,t,h}$ on a year counter variable by horizon sub-sample. This counter is set to zero before 1992 and increases by one every year after. We divide this variable by the number of years between 1993 and 2017 so that the estimated coefficient corresponds to the cumulative change in R^2 over the 1993-2017 period.[18] Results are reported in Table II and confirm the patterns in Figure IV. For horizons shorter than one and two years, the average R^2 has increased by 11.5 (Column (1)) and 9.4 percentage points (Column (2)), respectively. Beyond three and four years, the average R^2 has deteriorated by 11.5 (Column (4)) and 20 (Column (5)) percentage points. All four estimates are significant at the 1% level.

B The Slope of The Term Structure

To complement the previous approach, we study the evolution of the slope of the term structure, which we approximate every year by OLS. Figure V shows the year-on-year evolution of the slope estimates. The slope was around -10 until the mid-90s, but then became steeper every year (i.e., more negative), especially after 2005. After this date, the slope is consistently smaller than -10. Table III confirms this pattern. We regress the slope estimates on a year counter and a constant, as we did above. Column (1) shows that the term structure steepens over time, with an average slope that shifts from -6.6 during the baseline period 1983-1992 to (-6.6-10.6=) -17.2 in recent years.

[Insert Figure V and Table III about here]

2000 (compared to two years prior), whereas that of long-term forecasts declined.

[18] In this test and in the rest of the paper, we cluster standard errors by forecasted fiscal period, except in Table III where we cluster standard errors by year because observations are not available by forecasted fiscal period. In general, changing the level of clustering does not materially affect our statistical inference.

The rest of Table III shows that this pattern is robust to alternative estimation approaches. It holds in Columns (2) and (3) where we first estimate the slope by (two-digit SIC) industry and year; and in Columns (4) and (5) where we estimate it by analyst and year (for analyst-year with sufficient short and long-term forecasts). Results in Columns (3) and (5) are particularly remarkable. They indicate that the steepening of the term structure holds within industry and within analyst, and thus that the trend is not driven by changes in sample composition. Results in Column (5) also demonstrate that the selection of different analysts in our sample (some with, and others without missing long-term forecasts), cannot be the main explanation for our finding, since we observe the same change of the term structure for the same analyst over time.

C Robustness

The results of Table II and III survive many robustness tests, reported and discussed in Sections 9 and 10 of the online appendix. In brief, we find similar results when (i) controlling for the characteristics of the firms covered by the analyst, (ii) focusing on the sub-sample of analysts and firms with non-missing long-term forecasts, (iii) excluding the time periods with imperfect coverage by I/B/E/S such as the 80's, and (iv) using other periods than 1983-1992 as baseline. The results are also robust to alternative methodological choices to compute R^2.

In sum, the informativeness of analysts' short-term forecasts has improved over time while that of their long-term forecasts has declined. This pattern coincides with the rise of various sources of alternative data. Insofar as this evolution increases the marginal benefit of obtaining short-term information (our hypothesis), this aggregate evolution is consistent with our main prediction (Corollary 1).

VI Test#2: Exposure to Social Media Data

To complement the previous macro-level analysis, we now perform our second test at the micro-level exploiting variation of analysts' exposure to social media data generated on StockTwits.

A StockTwits Data

StockTwits was founded in 2008 as a networking platform for investors to share their opinions about stocks. Participants can post messages of up to 140 characters with extra content (e.g., charts, links) and make a buy ("Bullish") or sell ("Bearish") recommendation for the underlying stocks. They use $cashtags with stocks' ticker symbols to link their messages to firms. Users of StockTwits and its services include, for instance, retail investors, finance professionals (including analysts) and journalists.

We obtained data from StockTwits for all messages posted between January 1, 2009 and December 31, 2017. For each message, we observe the user identifier, the date, content, recommendation, and associated $cashtags with the corresponding tickers (a message can be associated with multiple tickers). We also have access to users' self-declared information, including their name and investment horizon, and for each firm, to its listing venue and its "watchlist", i.e., the number of users who explicitly follow that firm. We keep messages about firms trading on NASDAQ, NYSE, NYSEArca, NYSEMkt, or trading OTC, that are present in CRSP with share code 10, 11, and 12. These filters produce a sample of 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. Both have increased exponentially since 2009. The upper left panel indicates that the number of daily messages has increased from less than 1,000 to more than 80,000. The upper-right panel shows a similar trend in the average number of users on a firm's watchlist. The lower panels display the evolution of the distributions of both variables. They show substantial and increasing heterogeneity in activity across firms and time. This variation reflects the heterogeneous expansion of the platform, with some firms receiving high social media coverage early, some firms receiving coverage later, and others remaining outside most discussions.

B Relevance Conditions

Our test exploits this heterogeneous expansion across firms. We conjecture that analysts covering different sets of firms were differently exposed to new short-term oriented data produced by StockTwits. Our test thus requires that (i) discussions on StockTwits indeed contain information about firms' short-term prospects, and (ii) analysts use this information. Evidence from prior literature and our own tests support both conditions.

First, there is considerable evidence that social media data contain information about firms' future returns or earnings (Chen et al. (2014), Jame et al. (2016), Bartov et al. (2018), Tang (2018), Gu & Kurov (2020), or Leung et al. (2019)). However, evidence of earnings or sales predictability reported in these papers are typically for the current or next fiscal quarter.[19] We confirm that this predictive power vanishes at long horizons using "Bullish" and "Bearish" ratings issued by StockTwits' users. Specifically, we test whether these ratings predict firm growth at different horizons by estimating the following cross-sectional forecasting regression by quintile of total assets and year:

$$g_{j,y+h} = b_0 + b_1 Rating_{j,y} + b_2 g_{j,y-1} + \epsilon_{j,y}, \tag{16}$$

where j indexes all firms from the same fiscal year, y, and quintile of total assets. $Rating_{j,y}$ is the difference between the fractions of "Bullish" and "Bearish" messages about j over the current fiscal year y and $g_{j,y+h}$ is the future (year-on-year) growth observed in year $y+h$.[20] The main coefficient of interest is b_1. It measures the predicted change in growth in year $y+h$ associated with a change in rating today. Figure VII shows the means of b_1, by horizon $y+h$ when g is the growth of sales, EBITDA, EBIT, or Net Income. For all measures, better ratings predict higher growth but this association is statistically significant only at short

[19] This is consistent with anecdotal evidence from industry reports highlighting the short-term nature of social media data. For example, a brochure from Deutsche Bank emphasizes the usefulness of "Estimize" (a social media that crowdsources estimates of future earnings from many individuals) relative to other data sources. 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 (See "*The wisdom of crowds: crowdsourcing earnings estimates*", Deutsche Bank Market Research, March 4 2014).

[20] $Rating_{j,y}$ is not a text-based measure of sentiment. Unless missing, the "Bearish" or "Bullish" rating is publicly and directly observable without ambiguity.

horizons, i.e., for the current fiscal year $(y+0)$ and the next one $(y+1)$.[21]

[Insert Figure VII about here]

The second condition is that analysts use data from social media, including StockTwits. Section 13 of the online appendix provides an example of J.P.Morgan analysts using social media data. More systematic evidence is documented by Chi et al. (2021). They find that sentiment measures built from social media data come second in terms of the most frequently used alternative data by financial analysts, after app usage and on par with point-of-sale data. Among the different social media data providers, StockTwits is commonly referred to as a major one, especially for discussions about stocks.[22] StockTwits' datafeed has also been gradually integrated into all major financial information aggregation platforms used by practitioners (e.g., Bloomberg.com or Reuters.com), suggesting analysts are commonly exposed to this data.

We provide two additional sets of results indicating that analysts do rely on information produced on StockTwits. First we find that analysts are more likely to issue a new forecast on a given firm and day following an increase in StockTwits activity, including days without news arrival from traditional data sources (Table A1 in Section 6 of the online appendix). Moreover, we show that analysts are more likely to upgrade (downgrade) their recommendation for a stock when more users are "Bullish" ("Bearish") about a stock (Table A2 in Section 6 of the online appendix). Second, using biographic information (analysts' last names and the first letter of their first names) from I/B/E/S over the 2009-2017 period, we find

[21] Year-on-year growth for the current fiscal year $(g_{j,y+0})$ is known *after* the fiscal year is over, i.e., *after* observing the ratings issued *during* fiscal year y. Nonetheless, $Rating_{j,y}$ may reflect interim information publicly disclosed about $g_{j,y+0}$ around quarterly announcements. One way to solve this issue is to calculate *Rating* by fiscal quarter, and to do the same analysis with quarterly data. Doing so yields similar results.

[22] In their "2019 Alternative Data Handbook", J.P.Morgan analysts describe StockTwits as "*the leading (...) platform for the investing community (...), producing streams that are viewed by an audience of over 40 million across the financial web and social media platform*" (Source: J.P.Morgan (Oct. 25, 2019)). Commenting on the importance and visibility of StockTwits, other research analysts from Harbin write: "*Position-trading, day-trading, and swing-trading are now household terms for several million Americans, due to the skyrocketing number of part-time and full-time traders in the United States. As of this writing, the StockTwits platform boasts a rapidly-growing current user base of two million U.S.-based traders, which we believe to represent approximately one-third of active or semi-active traders in this country.*" (Source: Harbin Research (Sept. 14, 2020))

that 35% of (7,655 distinct) analysts' names exactly match those of StockTwits' account holders.[23]

C Test Specification

Our test compares 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" in which variation in data generated on StockTwits about a given firm due to StockTwits' staggered expansion is used to measure variations in 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. Our test begins on January 1, 2005 – almost 5 years before the first message we observe on the platform on July 13, 2009 – and ends on December 31, 2017. Our baseline specification by horizon sub-sample 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 η_t and η_i are time (i.e., date) 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). We further control for several characteristics of the portfolio of firms covered by the analyst at $t-1$.[24]

"Data Exposure" measures analyst i's exposure to social media data generated on 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 its expansion. We posit that analysts who are more exposed to StockTwits experience an increase in the volume of

[23] This finding is not evidence of analysts being active users. However, the mechanism we test only requires that analysts consume information from StockTwits, not that they communicate on this platform. Also, our matching analysis likely underestimates analysts' consumption of information on StockTwits because a StockTwits account is useful for receiving alerts on a specific list of stocks but not required for reading messages posted on the platform.

[24] Those characteristics are the mean characteristics of portfolio firms, and include size, age, cash flow to assets, cash to assets, debt to assets, and Tobin's Q. All explanatory variables in eq.(17) are standardized by their sample standard deviation, are winsorized at the 1% and 99% by date (t) (unless they are log-transformed variables), are measured at $t-1$, and are defined in Appendix II.

alternative social media data available to them. Accordingly, we predict 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).

We measure exposure to data generated on StockTwits in two distinct ways. One challenge is to ensure that these measures do not also capture how an analyst's exposure to data coming from *other* data sources concurrently changes with StockTwits' expansion. Messaging activity on StockTwits is indeed correlated with the arrival of information, whatever its origin. Hence the content on StockTwits does not only originate from the discussions on the platform but can 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 an analyst's exposure to data that is specifically generated on StockTwits and that would not (counterfactually) be available without this social media.

The first measure relies on the number of users who have on their "watchlist" the same firms as the ones covered by an analyst. 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 across the firms she covers. We then use this average number of users (denoted $\#Watchlist$) as a measure of her exposure to StockTwits data.[25] Importantly, a user's watchlist is persistent. This list is typically declared at the time of registration, and is rarely modified after. As a result, a firm's watchlist changes because new users register and enter the platform. Therefore, 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. As shown below, changes in a firm's watchlist are indeed largely uncorrelated with the arrival of information from traditional data sources (which could have affected analysts' forecasts in the absence of StockTwits).

The second measure relies on the volume of messages posted about the firms covered by an analyst. Because the number of actual messages may correlate with the arrival of information

[25]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 watchlist to zero.

from traditional data, we estimate *hypothetical* messages. We calculate every day the share of total messages posted on StockTwits about each firm j and compute the average share by firm, reflecting the usual daily share of all messages captured by j. We then multiply this average share by the total number of messages posted on StockTwits on day t, to obtain the number of messages for j that one would have expected to observe at t if the intensity of discussions about firm j (relative to the intensity of discussions about other firms) was at its average level. We sum the total number of such hypothetical messages in the last thirty days (from $t-30$ to $t-1$) for each firm, and take the average across the firms covered by each analyst (denoted *#Hypothetical Messages*). Importantly, the ratio of daily hypothetical messages over total actual messages is (by construction) constant within firm. Hence, the daily number of hypothetical messages *relative* to total actual messages varies across firms but not within. Since analyst coverage is persistent, most of this *relative* heterogeneity across firms is controlled for by the analyst fixed effects η_i in eq.(17). Consequently, the main source of variation used to isolate the effect of StockTwits is the *aggregate* number of actual messages, which is plausibly unrelated to individual firm and analyst characteristics, as well as the regular flow of firm-level information.[26]

Tables A4 and A5 (reported in Section 8 of the online appendix) present the results of two tests attempting to falsify our assumption that neither *#Watchlist* nor *#Hypothetical Messages* relates to the regular flow of firm-level information. We use Capital IQ Key Developments to identify firm-level news from traditional data sources and build two daily measures of news flow for a given firm: (i) the number of news events, and (ii) the total market response (in absolute value) to these news events to account for their relevance. Table A4 shows no significant relationship between the number of news events reported in Capital IQ and the number of (i) users in a firm's watchlist, or (ii) hypothetical messages. Table A5 shows similar results when using the market response to news arrival. Since we cannot observe all information, providing definitive evidence that our identifying assumption holds is not possible, but these two falsification tests demonstrate that it cannot easily be rejected.

[26]See Section 7 in the online appendix for more details on the sources of variations in *#Hypothetical Messages*.

[Insert Table IV about here]

Table IV presents summary statistics for the variables we use. The sample contains 31,623,819 daily observations over the 2005-2017 period. On average, R^2 is 68.33%, the forecasting horizon h is 1.26 years, and an analyst covers 10.35 firms. Each of these covered firms is typically followed by 321 users on StockTwits, and there are on average (138/30=)4.6 hypothetical messages about each, daily.

D The Effect of Exposure to StockTwits' Data

D.1 Forecast Informativeness by Horizon

Table V presents estimates of eq.(17) by horizon sub-samples using both measures of "Data Exposure". To ease economic interpretation, we normalize both variables by their sample standard deviation. Columns (1) and (2) show that increased exposure to StockTwits' data has a significantly *positive* effect on the informativeness of analysts' short-term forecasts ($h \leq 1$; horizon up to one year). In contrast, Columns (5) to (8) show that increased exposure to StockTwits has a significantly *negative* effect 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 results in a drop in R^2 between 1.51 and 1.92 percentage points for long-term forecasts, and an increase in R^2 between 0.53 and 0.66 for short-term forecasts. Columns (3) and (4) show no effect on informativeness for $1 < h \leq 2$, suggesting that the horizon of inflection, i.e., the value of h where the effect of greater exposure to StockTwits on R^2 changes sign, is between 1 and 2 years.

[Insert Table V about here]

D.2 The Slope of The Term Structure

Results from Table V are consistent with analysts allocating more (less) effort to the task of forecasting short-term(long-term) earnings, but do not explicitly control for uniform changes in informativeness common to *all* horizons. To better show that analysts substitute effort away from forecasting long-term earnings, we pool *all* sub-sample observations and modify eq.(17) to allow for an interaction term between "Data Exposure" and horizon h, which we

re-center at 1 and call h^* (i.e., $h^* = h - 1$). Specifically, we estimate:

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

In eq.(18), λ_0 measures the (unconditional) slope of the term structure, and λ_2 how it changes with greater data exposure, controlling for uniform changes in $R^2_{i,t,h}$ for all h which are captured by λ_1. Centering h at 1 is neutral on estimates for λ_2 (and λ_0), but it allows interpreting λ_1 as the effect of "Data Exposure" on R^2 at the 1-year horizon (and not zero), and thus to detect whether the term structure simply rotates ($\lambda_2 \neq 0$ and $\lambda_1 = 0$), or if it also shifts either upward ($\lambda_1 > 0$) or downward ($\lambda_1 < 0$). We re-center at 1 because Table V suggests the inflection horizon is between 1 and 2 years.

[Insert Table VI about here]

We report estimates of eq.(18) in Table VI for both measures of analysts' exposure. In Column (1), both λ_0 and λ_2 are negative and significant. The term structure is downward sloping ($\lambda_0 < 0$), and greater exposure to StockTwits makes it steeper ($\lambda_2 < 0$). Interestingly, λ_1 is not statistically different from zero, meaning that the slope of the term structure changes, but informativeness at the 1-year horizon does not. The term structure thus rotates around the one year horizon but does not experience a uniform shift upward or downward. This finding provides evidence of a "pure" reallocation effect of alternative data.

We obtain similar results when interacting h^* with the fixed effects (Column (2)), or when controlling for the characteristics of covered firms (Column (3)). Interacting h^* with the analyst fixed effects is equivalent to estimating λ_0 separately for each analyst, and thus allows controlling for permanent differences in the slope of the term structure across analysts. Likewise, interacting h^* with the date fixed effects allows controlling for the aggregate variations in the slope of the term structure, like the ones described in figure V, and that are unrelated to StockTwits expansion.[27] The rest of Table VI (Columns (4) to (6)) shows that our conclusions are similar when using *#Hypothetical Messages* to measure analysts' exposure.

[27]This full interaction approach is also used in papers that study difference-in-differences on a slope coefficient with fixed effects (see for instance eq.(8) in Edmans et al. (2017), and the discussion that follows).

D.3 Economic Magnitude

The estimate for λ_0 in the first column of Table VI implies that R^2 decreases by 16.66 percentage points for every one-year increase in horizon. This estimate differs from the estimate reported in Section IV.D because the sample period is more recent and the term structure has become steeper over time. Our estimate for λ_2 implies that a one standard deviation increase in exposure to StockTwits steepens (in absolute value) the slope of the term structure by 0.86, so that R^2 decreases by (16.66+0.86=)17.52 percentage points for every one-year increase in horizon.

The economic magnitude of this change in slope should be evaluated against normal variations for the slope of the term structure. Figure V displays yearly estimates at the aggregate level. The standard deviation of this time series is only 1.9 over the 2005-2015 period, but it is 4.5 when we consider the entire 1983-2015 period to obtain a more precise estimate. It is 8.6 when we estimate the slope by (2-digit SIC) industry and year from 2005 to 2015, and 11.1 when we do it by analyst and year over the same period. Therefore, the impact of analysts' exposure to StockTwits represents, on average, (0.86/4.5=)19.1%, (0.86/8.6=)10% and (0.86/11.1=)7.7% of the slope standard deviation at the aggregate, industry, and analyst-level, respectively. This economic magnitude is larger for analysts whose names match those of a StockTwits user account. For this sub-sample, and using the same specification as in Column (1) of Table VI (not reported for brevity), we find $\lambda_2 = -1.44$ (t-stat=-2.93), i.e., between 12.9% and 32% of the slope standard deviation.

D.4 Robustness

Our findings hold across several robustness tests, reported in Section 11 of the online appendix. In brief, results are similar when controlling for trading volume, and thus for the potential effect of news (public or private) that are material enough to generate trading. They are also robust to restricting our tests to (i) analysts who always cover the same firms, or (ii) analysts and firms with non-missing long-term forecasts. In sum, news arrival from other sources, changes in analyst coverage, or sample selection are unlikely to explain our findings.

E Additional Predictions and Ancillary Results

Results in the previous sections are consistent with our main prediction: as analysts become more exposed to alternative data, they substitute effort away from the task of forecasting long-term earnings at the benefit of the task of forecasting short-term earnings. According to our theory, this arises because alternative data raises the net marginal informational benefit of processing short-term information. To further support this mechanism, we test two ancillary predictions of our theory. Namely, the steepening of the term structure captured by λ_2 in eq.(18) should be more pronounced when (i) the cost of multi-tasking (c) is high, and (ii) earnings are less autocorrelated (so that the condition $\beta < (\frac{c}{2b})^{\frac{1}{2}} \frac{\psi_{lt}}{\psi_{st}}$ in Corollary 1 is more likely to be satisfied).

E.1 Multi-Tasking Costs (c)

First, we assess whether λ_2 is indeed more negative for analysts facing higher costs of multi-tasking. This should be the case for those who cover more stocks because the total number of forecasting tasks (within and across firms) increases with coverage.[28] We thus count the number of firms in analysts' portfolio and interact this variable (named $\#Firms$) with all variables in eq.(18) to test whether λ_2 is (even) more negative for analysts covering more firms. Results are in Table VII. The coefficients on the triple interaction term are consistently negative, indicating that λ_2 indeed decreases with $\#Firms$. As predicted, the steepening of the term structure is stronger with more costly multi-tasking. Two other coefficients are consistently negative and highly significant. The first is the coefficient on $\#Firms$, indicating that, everything else being equal, a greater multi-tasking cost negatively affects forecast informativeness for all h (which is consistent with the model). The second is the coefficient on $h^* \times \#Firms$, which shows that the tendency to allocate more effort to the task of forecasting short-term earnings when the number of forecasting tasks increases is also true unconditionally, i.e., with and without exposure to StockTwits. Put simply, the more tasks an analyst has, the more she focuses on the short-term.

[28] 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.

[Insert Table VII about here]

E.2 Correlated Earnings (β)

Our model also predicts that the steepening of the informativeness term structure due to greater exposure to alternative data should be more salient for analysts following firms whose long-term and short-term earnings are less correlated (β in the model). We estimate the correlation in earnings at the firm level by regressing firms' quarterly earnings on their lag (without constant) using a rolling window of two years (and requiring at least four observations). Then, we average the estimated autocorrelation across all firms covered by the analyst, and again interact this variable (named *Auto*) with all variables in eq.(18). Table VIII shows that the coefficients on the triple interaction term are all significantly positive. Thus, as predicted, the steepening of the term structure is less pronounced for analysts covering firms whose earnings are *more* autocorrelated. Again, the other regression coefficients are broadly consistent with our premises. For example, the coefficient on $h^* \times$ *Auto* is consistently positive and significant in four out of six specifications, implying that the slope of the term structure is usually less steep for analysts covering firms with greater earnings autocorrelation.

[Insert Table VIII about here]

F Alternative Explanations and Interpretations

Our findings are consistent with our prediction: greater exposure to alternative (social media) data increases the informativeness of short-term forecasts, but decreases that of long-term forecasts. Moreover, the heterogeneity of this effect across analysts can be explained by our theory. Other factors influencing R^2 may arguably explain our findings. These factors can be broadly classified into three main categories.

The first category includes variables related to public information. One concern (discussed above) is indeed that our findings reflect changes in information available to analysts from sources other than StockTwits. However, Tables A4 and A5 in Section 8 of the online appendix rule out this possibility. The second category includes variables related to earnings

uncertainty. When uncertainty is higher, forecasts are naturally less precise, and thus less informative, so our results may arise because uncertainty about earnings changes concurrently with the expansion of StockTwits. However, R^2 is explicitly designed to isolate forecasting informativeness from forecasting difficulty. The third category includes all variables affecting γ (i.e., analysts' horizon-specific incentives), including compensation schemes, career prospects, investors' demand, or brokers' internal organization. Nevertheless, to explain our results, changes in these variables should simultaneous trigger (i) an increase in R^2 for low values of h, (ii) a decrease in R^2 for high values of h, but (iii) no change in average R^2 across h (i.e., not only $\lambda_2 < 0$, but also $\lambda_1 = 0$ in eq.(18)). Moreover, they should systematically coincide with the timing of StockTwits' expansion across stocks, as well as the aggregate variation in messaging and following 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 interpretation of our findings could be that the introduction of StockTwits influences analysts' allocation of effort because StockTwits helps them to learn about investors' demand for short and long-term information. However, this learning channel requires that the clients of the analysts' employers and the individuals active on StockTwits are the same investors, or at least that there is sufficient overlap between them such that analysts can actually learn about information demand from StockTwits activity. This overlap is not present in our setting: brokerage house customers are mostly institutional investors, whereas StockTwits users are mainly retail. Moreover, for analysts covering a fixed portfolio of firms, and for whom incentives are stable, learning about demand is less likely to play some role, and yet Table A9 in Section 11 of the online appendix shows that our results also hold for those analysts.

VII Conclusion

This paper examines the effect of alternative data on the informativeness of financial forecasts. We posit that alternative data reduces the cost of producing information about short-term cash flows relatively more than that about long-term cash flows. We show theoretically

that this effect can induce forecasters of firms' cash flows to allocate more effort to the collection of short-term information at the expense of the collection of long-term information (as they optimally equalize the marginal net benefit of each type of effort). As a result, the informativeness of their forecasts of short-term cash flows improves while the informativeness of their forecasts about long-term cash flows can drop. Our main contribution is to test this novel prediction and confirm it. Using a large sample of short and long-term forecasts issued by sell-side equity analysts, we find that increases in the availability of alternative data (their overall expansion over time and the expansion of social media data from StockTwits) are associated with a drop in the informativeness of analysts' long-term earnings forecasts (more than two years), even though the informativeness of their short-term (less than one year) forecasts improves.

References

Abis, S. (2018), Man vs machine: Quantitative and discretionary equity management, Working paper, Columbia University.

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. (2018), 'Can Twitter help predict firm-level earnings and stock returns?', *The Accounting Review* **93**, 25–57.

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

Bloom, N. (2014), 'Fluctuations in uncertainty', *Journal of Economic Perspectives* **28**, 153–176.

Bradshaw, M. T. (2004), 'How do analysts use their earnings forecasts in generating stock recommendations?', *The Accounting Review* **79**, 25–50.

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.

Chen, L., Da, Z. & Zhao, X. (2013), 'What drives stock price movements?', *Review of Financial Studies* **26**, 841–876.

Chi, F., Hwang, B. & Zheng, Y. (2021), The use and usefulness of big data in finance: Evidence from financial analysts, Working paper, Cornell University.

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

Cookson, T., Engelberg, J. & Mullins, W. (2020*b*), Echo chambers, Working paper, University of Colorado at Boulder.

Copeland, T., Dogloff, A. & Moel, A. (2004), 'The role of expectations in explaining the cross-section of stock returns', *Review of Accounting Studies* **9**, 149–188.

Da, Z. & Warachka, M. (2011), 'The disparity between long-term and short-term forecasted earnings growth', *Journal of Financial Economics* **100**, 424–442.

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

Edmans, A., Jayaraman, S. & Schneemeier, J. (2017), 'The source of information in prices and investment-price sensitivity', *Journal of Financial Economics* **126**, 74–96.

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

Farboodi, M. & Veldkamp, L. (2020), 'Long run growth of financial data technology', *American Economic Review* **110**, 2485–2523.

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* **33**, 1367–1411.

Gerken, W. & Painter, M. (2021), The value of differing points of view: Evidence from financial analysts' geographic diversity, Technical report.

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**, 3–43.

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

Green, T. C., Huang, R., Wen, Q. & Zhou, D. (2019), 'Crowdsourced employer reviews and stock returns', *Journal of Financial Economics* **134**, 236–251.

Grennan, J. & Michaely, R. (2020*a*), Artificial intelligence and the future of work: Evidence from analysts, Working paper, Duke University.

Grennan, J. & Michaely, R. (2020*b*), 'FinTechs and the market for financial analysis', *Journal of Financial and Quantitative Analysis* p. 1–31.

Gu, C. & Kurov, A. (2020), 'Informational role of social media: Evidence from twitter sentiment', *Journal of Banking and Finance (forthcoming)* .

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* **133**, 83–98.

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

Huang, J. (2018), 'The customer knows best: The investment value of consumer opinions', *Journal of Financial Economics* **128**, 164–182.

Huang, S., Xiong, Y. & Yang, L. (2020), Skills acquisition and data sales, Working paper, University of Toronto.

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

Jung, B., Shane, P. & Yang, Y. (2012), 'Do analysts long-term growth forecasts matter? Evidence from stock recommendations and career outcomes', *Journal of Accounting and Economics* **53**, 55–76.

Katona, Z., Painter, M., Patatoukas, P. N. & Zeng, J. (2021), On the capital market consequences of alternative data: Evidence from outer space, Working paper, Saint Louis University.

Leung, W., Wong, G. & Wong, W. (2019), Social-media sentiment, portfolio complexity, and stock returns, Working paper, University of Edinburgh.

Martin, I. & Nagel, S. (2020), Market efficiency in the age of big data, Working paper, NBER.

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**, 653–686.

Mest, D. P. & Plummer, E. (1999), 'Transitory 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.

Mihet, R. (2020), Financial innovation and the inequality gap, Working paper, University of Lausanne.

Mukherjee, A., Panayotov, G. & Shon, J. (2021), 'Eye in the sky: Private satellites and government macro data', *Journal of Financial Economics* **141**, 234–254.

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.

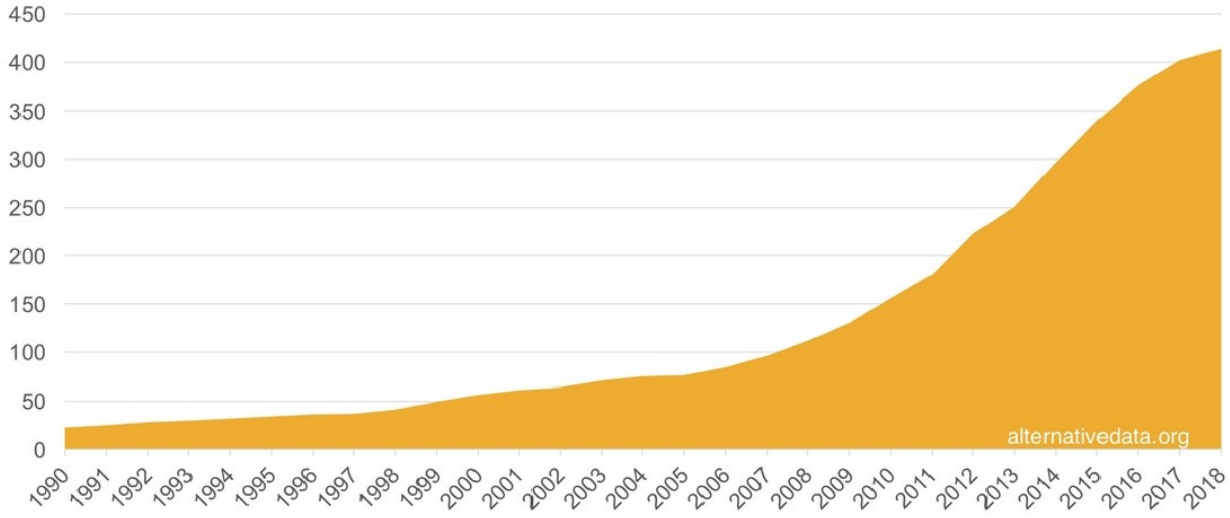
Tang, V. W. (2018), 'Wisdom of crowds: Cross-sectional variation in the informativeness of third-party-generated product information on Twitter', *Journal of Accounting Research* **56**, 989–1034.

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

Verrecchia, R. (1982), 'Information acquisition in a noisy rational expectations economy', *Econometrica* **50**, 1415–1430.

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

Figure I: Number of alternative data providers by year



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 II: Timeline of the model

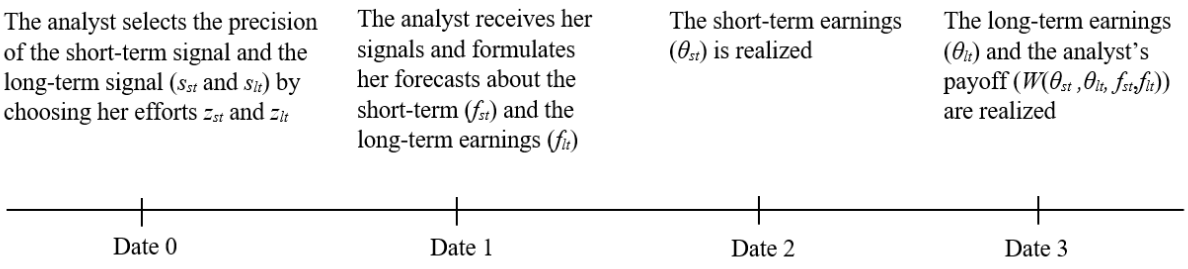
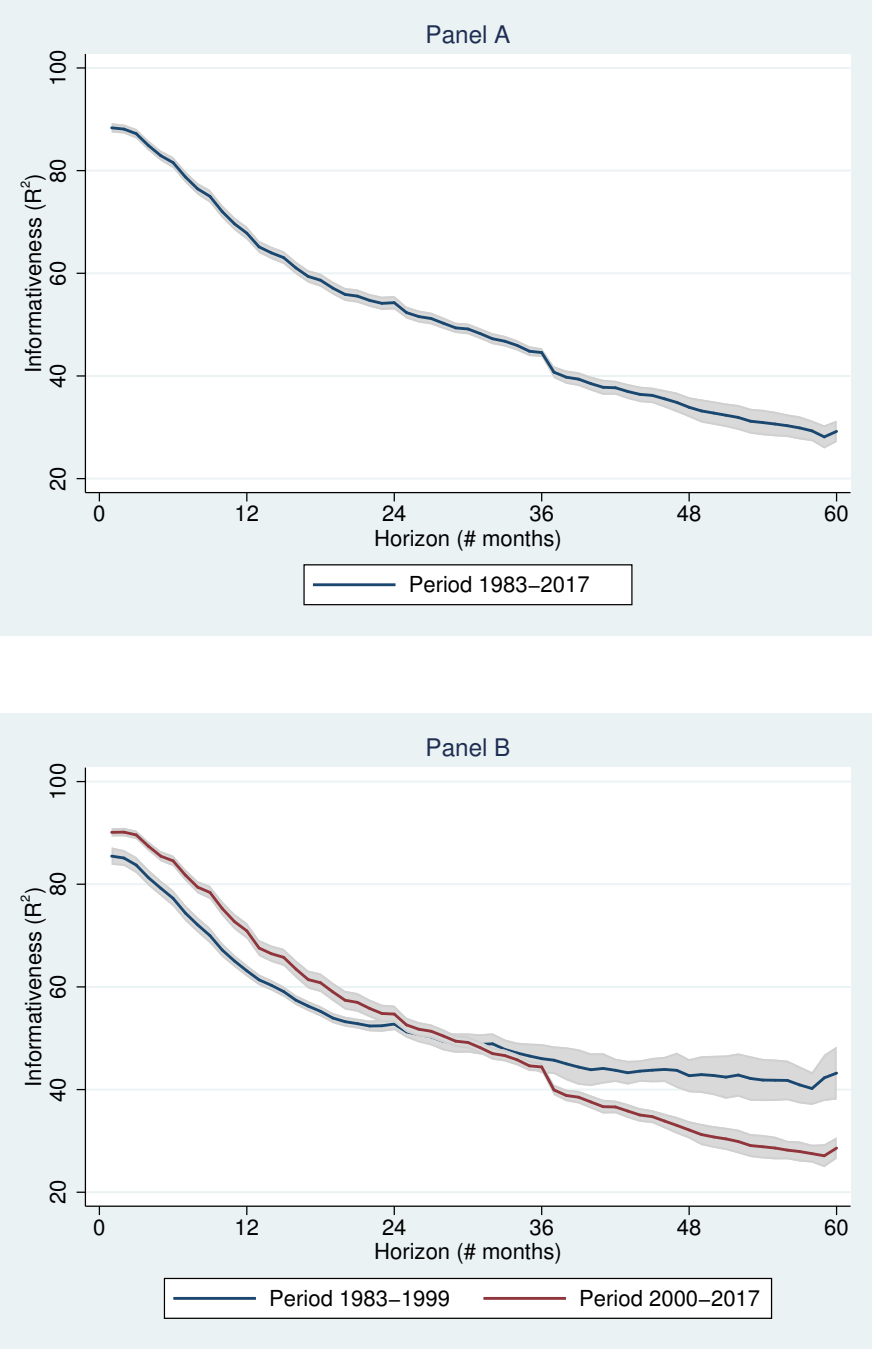
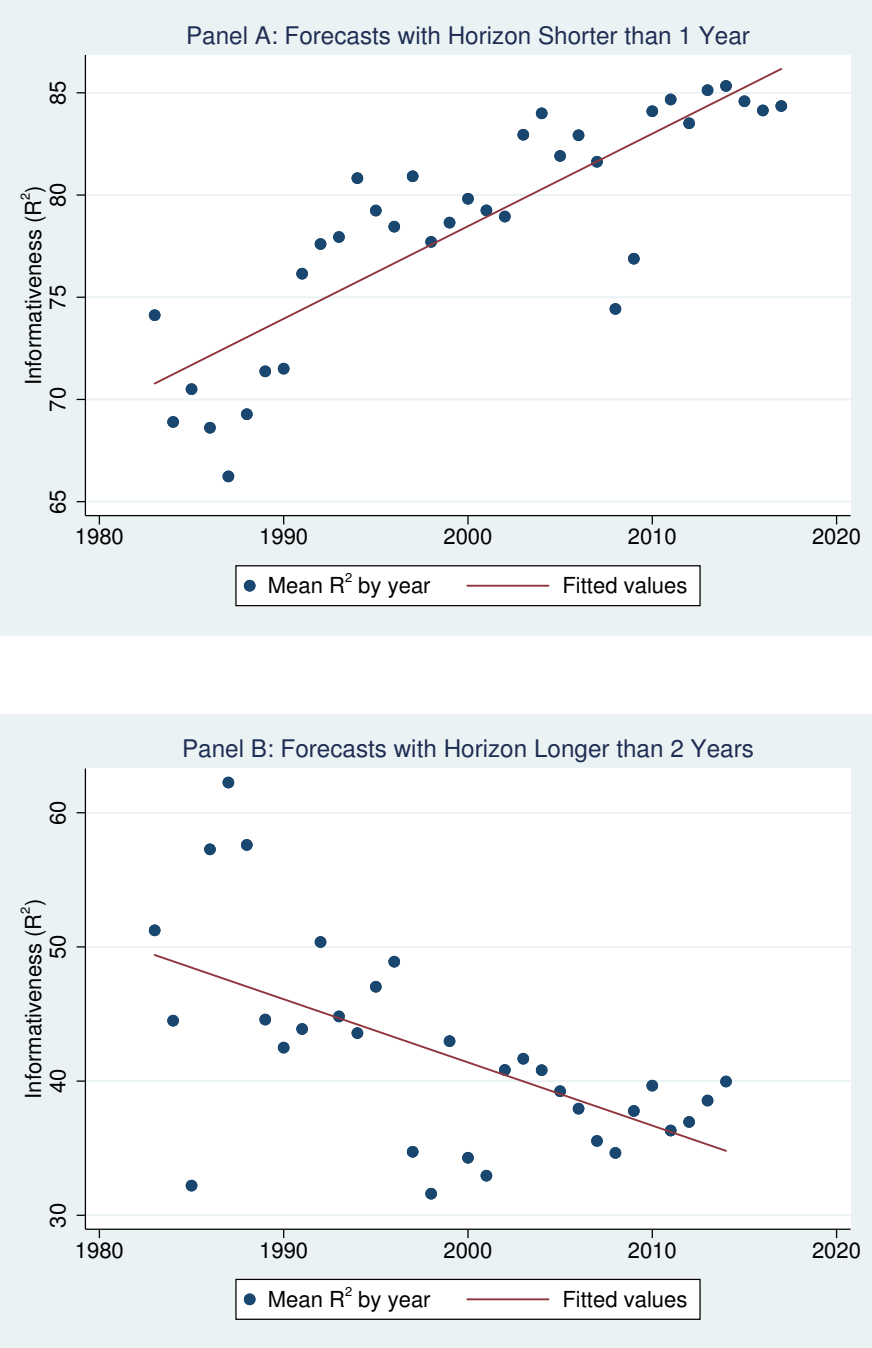


Figure III: The term structure of analysts' forecast informativeness



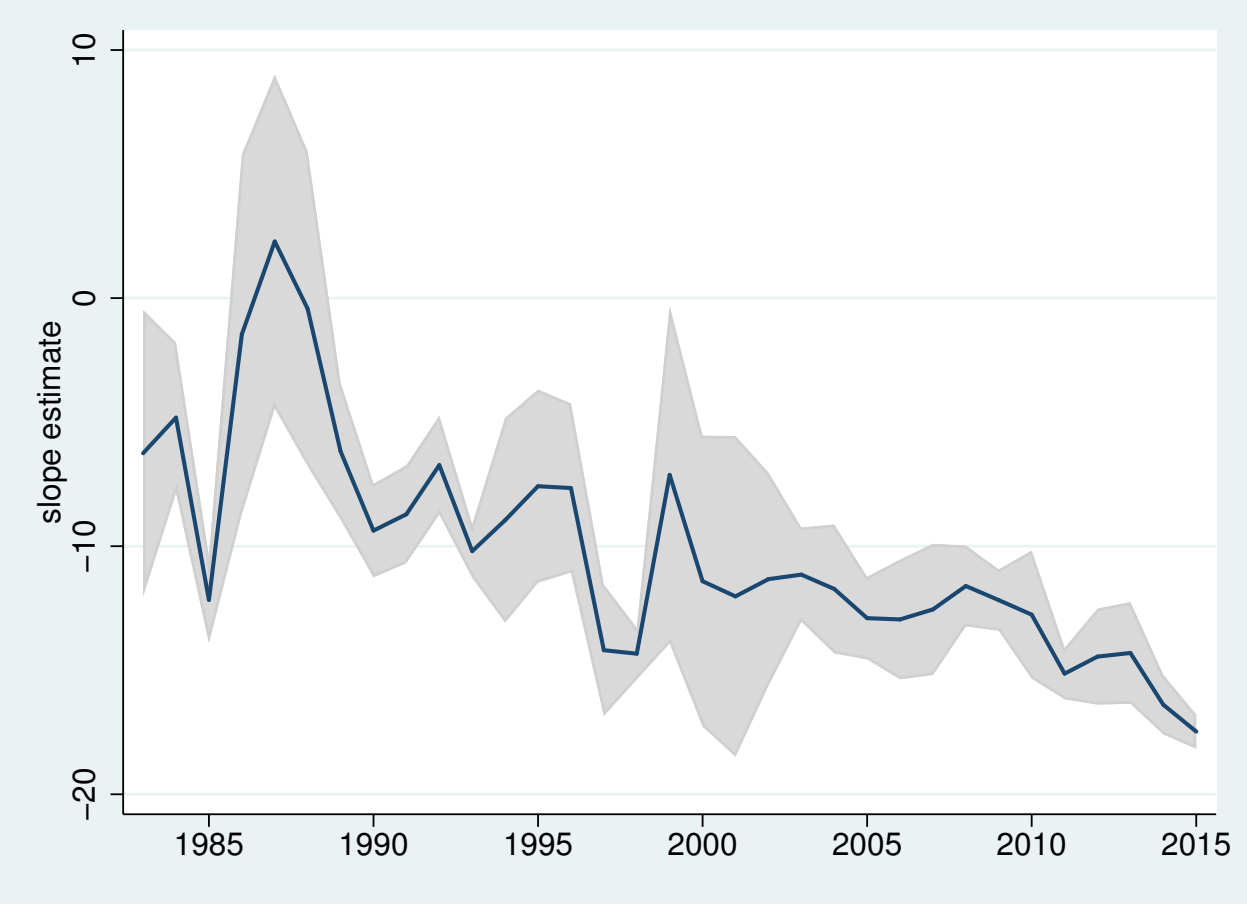
This figure displays the term structure of analysts' forecast informativeness. Each graph shows the means of analyst-level $R^2_{i,t,h}$ over all i and t, for fixed h values expressed in number of months (displayed on the x-axis). The forecasting horizon h is measured as the number of days between the forecasting date and the date of actual earnings release, divided by 365. The sample period is 1983-2017 (Panel A), split into two sub-periods (Panel B). The shaded gray area corresponds to a 90% confidence interval.

Figure IV: Short vs. long-term forecast informativeness by year



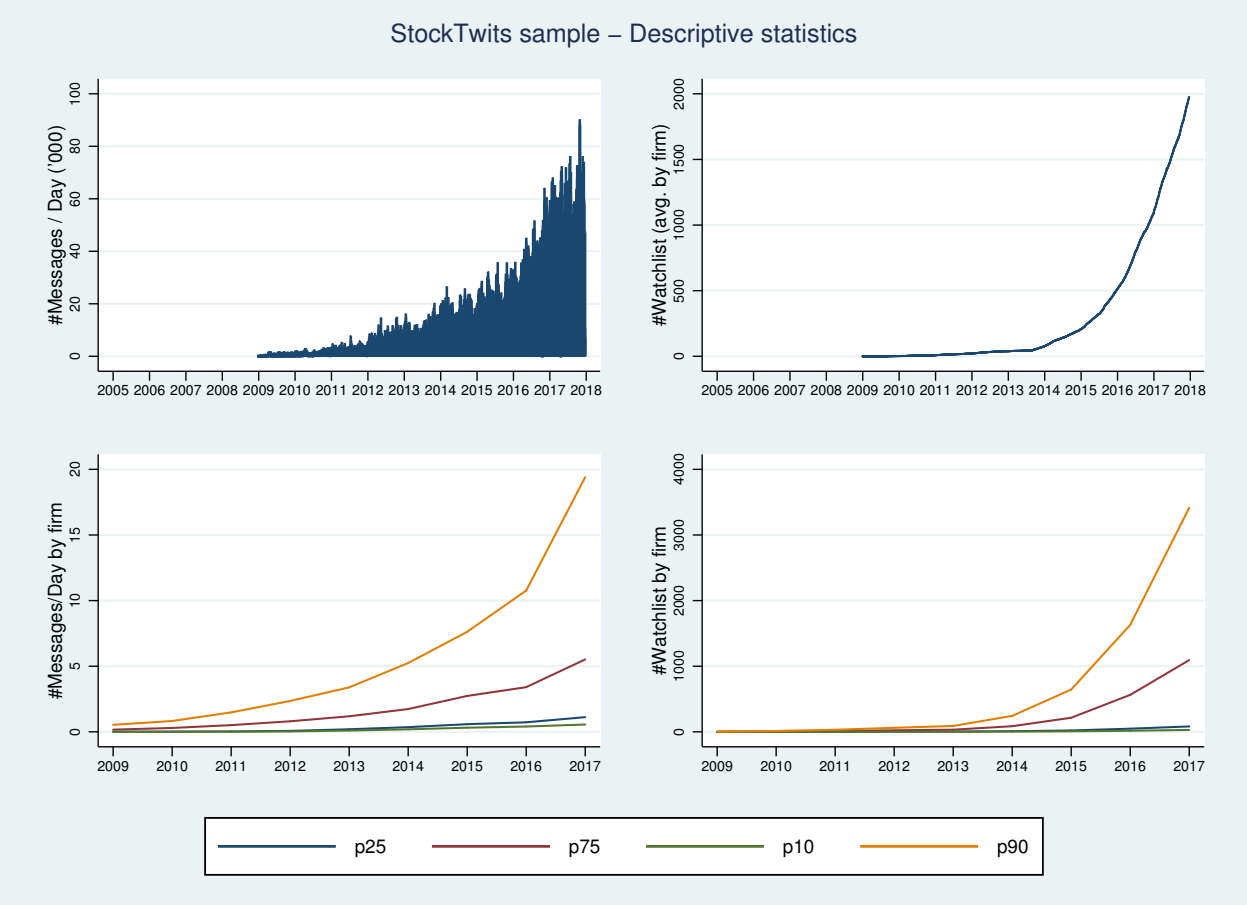
This figure shows the means of analyst-level $R^2_{i,t,h}$ over all i by year, separately for short ($h < 1$) and long-term ($h >= 2$) forecasts. The sample period is 1983-2017 for short-term forecasts (Panel A), and 1983-2015 for long-term forecasts (Panel B).

Figure V: The slope of term structure by year



This figure shows the evolution over time of the slope of the term structure of analysts forecasts' informativeness. The slope of the term structure is estimated every year by linear approximation. We do so by regressing the average of R^2 by horizon h on h, separately for every calendar year. R^2 measures analysts forecasts' informativeness. h measures the forecasting horizon (defined as the number of days between the forecasting date and the date of actual earnings release divided by 365). The figure plots the regression coefficient on the regressor h, which is an estimate of the slope of the term structure. Each slope estimate measures how R^2 changes in percentage points for every annual increment of h. For example, a slope estimate of -10 in 1993 indicates that in 1993, R^2 decreases on average by 10 percentage points when h increases by 1 year. The shaded gray area corresponds to a 90% confidence interval.

Figure VI: StockTwits' expansion



This figure shows 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. The bottom-left panel presents different percentiles of the number of messages per day and firm. The bottom-right panel presents different percentiles of the number of users that have a given firm in their watchlist.

Figure VII: StockTwits ratings and firm growth predictability

This figure shows the predictive power of Buy ("Bullish") and Sell ("Bearish") ratings issued by StockTwits users about firm growth, by horizon. It relies on cross-sectional forecasting regressions, estimated on every fiscal year starting from 2010 by quintile of total assets, and specified as follows:

$$g_{j,y+h} = b_0 + b_1 Rating_{j,y} + b_2 g_{j,y-1} + \epsilon_{j,y}$$

where j indexes all firms from the same quintile and year. $Rating_{j,y}$ is the difference between the fraction of "Bullish" messages and that of "Bearish" messages about firm j during fiscal year y, and $g_{j,y+h}$ is the (year-on-year) growth reported ex-post, in fiscal year $y+h$. $Rating_{j,y}$ is naturally bounded between -1 (all ratings issued during fiscal year y are "Bearish") and +1 (all ratings issued during fiscal year y are "Bullish"). The figure shows the means of b_1 (with the associated 90% confidence interval) for all quintiles and years y, by horizon $y+h$ (displayed on the x-axis) when g is the growth of sales (upper-left), EBITDA (upper-right), EBIT (bottom-left), or Net Income (bottom-right). The average predicted change in growth (in percentage points) associated with a change in rating today is displayed on the y-axis.

Table I: R^2 measure summary 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. 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. #Firms is the number of firms the analyst covers. 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 annual forecasting horizon ranges. Variable definitions are in Appendix II.

	N	Mean	St.Dev	Min	P25	P50	P75	Max
Whole sample								
R^2	65,889,122	68.01	33.90	0.00	45.71	82.70	96.30	100.00
h	65,889,122	1.11	0.83	0.00	0.48	0.99	1.56	5.00
#Firms	65,889,122	8.12	5.18	3.00	4.00	7.00	11.00	30.00
Sample: $0 < h \leq 1$								
R^2	33,413,667	79.60	27.63	0.00	72.57	92.49	98.42	100.00
h	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: $1 < h \leq 2$								
R^2	25,060,925	59.21	34.64	0.00	29.37	69.51	90.42	100.00
h	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: $2 < h \leq 3$								
R^2	5,361,069	49.37	36.23	0.00	10.47	53.15	84.34	100.00
h	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: $3 < h \leq 4$								
R^2	1,349,749	37.62	36.04	0.00	0.00	28.84	71.60	100.00
h	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: $4 < h \leq 5$								
R^2	703,712	31.18	34.98	0.00	0.00	14.75	62.31	100.00
h	703,712	4.43	0.28	4.00	4.19	4.40	4.65	5.00
#Firms	703,712	6.26	3.54	3.00	4.00	5.00	8.00	30.00

Table II: Forecast informativeness by horizon

This table presents OLS estimates of time trend in analysts' forecast 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. 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 cumulative increment in R^2 over the 1993-2017 period. Variable definitions are in Appendix II. 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:	$0 < h \leq 1$	$1 < h \leq 2$	$2 < h \leq 3$	$3 < h \leq 4$	$4 < h \leq 5$
OLS:	(1)	(2)	(3)	(4)	(5)
Year Trend	11.5***	9.4***	2.4	-11.5***	-20.0***
	(8.00)	(6.89)	(1.46)	(-5.12)	(-5.41)
Constant (83-92)	74.7***	55.0***	47.9***	44.3***	42.6***
	(93.81)	(82.46)	(39.10)	(29.78)	(21.12)
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712

Table III: The slope of the term structure

This table presents OLS estimates of time trend in the slope of the term structure. The dependent variable is the slope of the term structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. A negative slope indicates that forecast informativeness (R^2) 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. In columns (2) to (5), the regression coefficients on h used as estimates for the slope are winsorized by year at the 1% level in each tail. 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 cumulative change in slope over the 1993-2017 period. 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.

Dep. variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS:	(1)	(2)	(3)	(4)	(5)
Year Trend	-10.8***	-4.9***	-3.4***	-4.9***	-2.7**
	(-6.74)	(-4.75)	(-3.03)	(-7.60)	(-2.07)
Constant (83-92)	-6.5***	-11.3***		-12.1***	
	(-6.45)	(-20.53)		(-25.95)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	7,657	7,290

Table IV: StockTwits sample descriptive statistics

This table presents descriptive statistics for the main analyst-day-horizon variables in the StockTwits sample. The sample covers the period 2005-2017. R^2 measures the informativeness of the forecasts made by an analyst on a given day for a given 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. #Firms is the number of firms that the analyst covers. #Watchlist is the average number of users that have in their watchlist the firms covered by an analyst. 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 the firms (in the last thirty days) that an analyst covers. Both variables are set to zero prior to StockTwits' introduction in 2009. Auto is the average earnings autocorrelation across the firms covered by an analyst. The other variables are control variables used in the analysis. Detailed variable definitions are provided in Appendix II.

	N	Mean	STDV	Min	P25	P50	P75	Max
R^2	31,623,819	68.33	33.76	0.00	46.43	83.10	96.36	100.00
h	31,623,819	1.26	0.93	0.00	0.54	1.11	1.77	5.00
#Firms	31,623,819	10.35	5.40	3.00	6.00	9.00	13.00	29.00
#Watchlist	30,959,282	321	1,471	0	0	12	117	44,145
#Messages	30,959,282	112	413	0	0	16	76	13,044
#Hypothetical Messages	30,959,282	138	486	0	0	19	99	13,322
Auto	29,364,951	0.67	0.21	-0.01	0.55	0.69	0.82	1.12
Total assets	29,391,344	11,738	32,854	0	1,548	4,616	12,635	2,087,821
Total assets (Log)	29,391,344	8.35	1.54	-4.65	7.34	8.44	9.44	14.55
Age	29,392,961	22.97	12.41	1.00	13.43	20.24	29.90	68.00
Age (Log)	29,392,961	2.98	0.57	0.00	2.60	3.01	3.40	4.22
Cash flow to assets	29,384,430	0.05	0.12	-0.68	0.04	0.08	0.11	0.24
Cash to assets	29,391,077	0.21	0.17	0.01	0.08	0.15	0.30	0.88
Debt to assets	29,391,344	0.24	0.14	0.00	0.13	0.22	0.32	0.85
Tobin's Q	29,366,671	2.29	1.05	0.71	1.54	2.00	2.74	7.34

Table V: Data exposure and forecast informativeness by horizon

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) at different horizons to analysts' exposure to social media data generated on 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. We pool horizons between three and five years because we have few observations at long horizons. 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 a variable capturing the 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 forecast 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 firms covered by the analyst in their watchlist. In Panel B, Data Exposure is based on the number of hypothetical messages posted about the firms covered by the analyst 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 Appendix II. 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:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$h > 3$	
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Proxy for Data Exposure = #Watchlist							
Data Exposure	0.54***	0.53***	0.40	0.17	-0.66***	-1.01***	-1.51***	-1.55***
	(3.89)	(4.03)	(1.06)	(0.47)	(-3.24)	(-4.80)	(-3.49)	(-3.20)
Analyst 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	14,055,963	13,033,456	11,489,986	10,601,113	3,916,280	3,636,242	1,496,954	1,435,797
	Panel B: Proxy for Data Exposure = #Hypothetical Messages							
Data Exposure	0.66***	0.61***	0.56	0.16	-0.60*	-1.23***	-1.84***	-1.92***
	(4.39)	(4.46)	(1.27)	(0.36)	(-1.63)	(-4.17)	(-3.95)	(-3.43)
Analyst 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	14,055,963	13,033,456	11,489,986	10,601,113	3,916,280	3,636,242	1,496,954	1,435,797

Table VI: Data exposure and the slope of term structure

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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 Proxy:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-0.86***	-0.78***	-0.96***	-0.69***	-0.94***	-1.05***
	(-2.59)	(-3.06)	(-3.72)	(-2.75)	(-4.54)	(-5.03)
Data Exposure	0.13	-0.17	-0.35	0.34	-0.14	-0.32
	(0.50)	(-0.64)	(-1.29)	(1.42)	(-0.57)	(-1.30)
h^*	-16.66***			-16.62***		
	(-33.85)			(-32.13)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,281	30,105,556	27,860,429	30,959,281	30,105,556	27,860,429

Table VII: Differential effects by analysts' processing constraints

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) 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. Data Exposure is a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). #Firms is the number of firms that the analyst covers. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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)
$h^*\times$ Data Exposure $\times$ $\#Firms$	-0.14***	-0.06***	-0.06***	-0.10***	-0.04*	-0.06***
	(-5.71)	(-3.39)	(-3.82)	(-6.18)	(-1.64)	(-2.56)
$h^*\times$ Data Exposure	0.69	-0.04	-0.23	-0.06***	-0.03	-0.03
	(1.61)	(-0.10)	(-0.74)	(-2.58)	(-0.99)	(-1.25)
$h^*\times$ $\#Firms$	-0.15***	-0.23***	-0.23***	-0.14***	-0.23***	-0.22***
	(-6.58)	(-8.67)	(-8.24)	(-5.96)	(-8.63)	(-8.01)
Data Exposure $\times$ $\#Firms$	-0.09***	-0.05***	-0.04**	-0.06***	-0.03	-0.03
	(-3.34)	(-2.88)	(-2.26)	(-2.58)	(-0.99)	(-1.25)
$\#Firms$	-0.22***	-0.23***	-0.25***	-0.23***	-0.24***	-0.25***
	(-5.97)	(-6.95)	(-7.00)	(-5.79)	(-6.99)	(-6.92)
Data Exposure	1.10***	0.42	0.14	0.98***	0.16	0.01
	(2.80)	(1.48)	(0.53)	(3.16)	(0.46)	(0.02)
h^*	-15.00***			-15.05***		
	(-23.62)			(-22.82)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,281	30,105,556	27,860,429	30,959,281	30,105,556	27,860,429

Table VIII: Differential effects by earnings' autocorrelation

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Auto is the average earnings' autocorrelation in analysts' portfolios. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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)
$h^*\times$ Data Exposure $\times$ Auto	1.17***	0.64***	0.58***	0.69***	0.39**	0.35**
	(3.23)	(2.82)	(2.62)	(2.75)	(2.18)	(2.00)
$h^*\times$ Data Exposure	-4.85***	-2.92***	-2.83***	-3.19***	-2.21***	-2.15***
	(-3.88)	(-4.34)	(-4.28)	(-3.52)	(-4.03)	(-4.14)
$h^*\times$ Auto	0.62**	0.57***	0.55***	0.35*	0.1	0.14
	(1.95)	(3.07)	(3.12)	(1.77)	(0.66)	(0.90)
Data Exposure $\times$ Auto	1.17***	0.64***	0.58***	0.39**	0.40***	0.39***
	(3.23)	(2.82)	(2.62)	(1.99)	(2.92)	(2.88)
Auto	1.68***	1.74***	1.32***	1.68***	1.73***	1.31***
	(7.38)	(8.88)	(6.66)	(7.23)	(8.63)	(6.43)
Data Exposure	-2.02*	-2.14***	-2.22***	-1.14	-1.52***	-1.66***
	(-1.80)	(-3.25)	(-3.33)	(-1.61)	(-2.71)	(-3.05)
h^*	-18.07***			-17.91***		
	(-28.26)			(-25.76)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	28,712,339	27,865,920	27,840,983	28,712,339	27,865,920	27,840,983

VIII Appendix

Appendix I – R^2 Estimation Procedure

This appendix shows how to estimate our measure of analysts' forecasts' informativeness, R^2, based on the initial sample of 9,129,282 unique forecasts and realizations described in Section IV.A. We illustrate our procedure with a fictitious analyst XYZ covering 6 firms (A, B, C, D, E, F) on January 19, 2007, and making earnings forecasts for the fiscal period ending December 31, 2008. The procedure consists of five steps:

- Step 1: Identify the future fiscal period of interest. Analysts make separate forecasts for the current fiscal period, for the next fiscal period, and for the subsequent ones. 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. Her forecast is then regarded as stale and the R^2 measure is computed excluding the underlying stock.[29] Column 1 of Table IX below shows the last available earnings forecasts made by XYZ for A, B, C, D, E, and F as of January 19, 2007. The actual realized earnings for fiscal year 2008 are in Column 2.[30]

- Step 3: Normalize earnings. Heterogeneity across firms on size is persistent. To exclude this persistent size effect from our R^2 measure, we normalize both earnings forecasts and realized earnings for each firm by its total assets at the end of the forecasted fiscal period. Total assets as of December 31, 2008 for A, B, C, D, E and F are in Table IX, Column 3. Earnings forecasts ($\widehat{e}_j$) and realized earnings (e_j) after normalization are reported in Columns 5 and 6.[31]

- Step 4: Estimate eq.(14) by OLS and compute R^2. Regress e_j on $\widehat{e}_j$ in the cross-section of covered firms j (i.e., across A, B, C, D, E and F) and calculate the R^2 of the

[29] For example, if as of January 19, 2007, 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.

[30] Notice Step 2 assumes that the belief of XYZ about an individual firm does not change until a new forecast is disclosed. Our results are similar if we relax this assumption by estimating first all unobserved forecasts between two consecutive observable forecasts by linear interpolation, and then use these interpolated forecasts instead of the last available forecast to compute R^2 daily.

[31] Our results are robust to different normalization approaches. In this example, total assets could be measured as of December 31, 2006, i.e., from the last available financial statements on January 19, 2007. One drawback with this alternative approach is that the measure of informativeness will change even when analysts do not update their forecasts (because the normalization changes).

regression. R^2 is set to zero if $\widehat{e}_j$ negatively predicts e_j (i.e., if $k_1 < 0$ in eq.(14)). It is set to missing if there are fewer than 3 or more than 30 observations in the regression, or if k_1 is missing after trimming that regression coefficient at the 1% level in each tail.[32] In Table IX, the R^2 of the regression of e_j (Column 6) on $\widehat{e}_j$ (Column 5) for XYZ on January 19, 2007 is 14.9%.

- Step 5: Compute the horizon. Horizon is the time elapsed until actual earnings are reported. Since earnings report 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. Column 4 from Table IX shows that realized earnings for A, B, C, D, E, and F, were all reported on March 31, 2009, so the median date is March 31, 2009. The horizon associated with the above R^2 of 14.9% is thus 2.20 years (802 days, divided by 365).

Table IX: Example of R^2 computation for analyst XYZ on January 19, 2007

Forecasted Fiscal Period: 12/31/2008

Firm	Latest Forecast ($M) (1)	Realized Earnings ($M) (2)	Total Assets ($M) (3)	Earnings Report Date (4)	Latest Normalized Forecast ($\widehat{e}_j$) (5)	Realized Normalized Earnings (e_j) (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

At the end of the above procedure, we find $R^2_{i,t,h} = 14.9\%$ for $i =$"XYZ", $t =$"January 19, 2007", and $h =$2.20. We apply the same procedure every day from January 1, 1983 to December 31, 2017 to every analyst in our sample for all available forecasted fiscal periods. This procedure yields a sample of 65,889,122 daily observations of R^2 with an associated horizon between 1 day and 5 years across 14,379 distinct analysts.

[32]Trimming of k_1 is possible ex-post, after all observations of R^2 are available. This filter reduces the effect of outliers coming from lower power in estimations of eq.(14) with few observations.

Appendix II – 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.
h	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.
h^*	Horizon h centered at 1 ($h^* = h - 1$)
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 $\overline{w_j} \times N_t$, where $\overline{w_j}$ is the mean of $w_{j,t}$ for all t after a message is observed for the first time, and N_t is the total number of messages posted about all firms at time t. $w_{j,t}$ is defined as $\frac{\#Messages_{j,t}}{N_t}$.
#Watchlist	Total number of StockTwits users having a given firm in their watchlist.
Age	1+number of years in Compustat since inception.
Auto	Within firm quarterly net income (*ibq* item in Compustat) autocorrelation, 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).
Total assets	at (from last available financial statements in Compustat).
Trading volume	Total number of shares traded from $t-30$ to $t-1$.

Appendix III – Derivations in the Model

Proof of Equation (5). Differentiating $\overline{W}(f_{st}, f_{lt}; s_{st}, s_{lt})$ with respect to f_{st} and f_{lt}, we obtain that the first order conditions to the analyst's problem at date 1 are:

$$\begin{aligned} \frac{\partial \overline{W}}{\partial f_{st}} &= -2\gamma(f^*_{st} - \mathsf{E}(\theta_{st} \mid s_{st}, s_{lt})) = 0 \\ \frac{\partial \overline{W}}{\partial f_{lt}} &= -2(1-\gamma)(f^*_{lt} - \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt})) = 0. \end{aligned} \tag{19}$$

Solving for f^*_{st} and f^*_{lt} and using the fact that s_{lt} is uninformative about θ_{st}, we obtain eq.(5). It is straightforward that the second order conditions are satisfied.

Proof of Equation (6). Substituting eq.(5) into eq.(4), we have:

$$\begin{aligned} \mathsf{E}(\overline{W}(f^*_{st}, f^*_{lt}; s_{st}, s_{lt})) &= \omega - \gamma\,\mathsf{E}((\mathsf{E}(\theta_{st}\,|s_{st}) - \theta_{st})^2) - (1-\gamma)\,\mathsf{E}((\mathsf{E}(\theta_{lt}\,|s_{st}, s_{lt}) - \theta_{lt})^2), \\ &= \omega - \gamma\,\mathsf{E}(\mathsf{Var}(\theta_{st}\,|s_{st})) - (1-\gamma)\,\mathsf{E}(\mathsf{Var}(\theta_{lt}\,|s_{lt}, s_{st})), \\ &= \omega - q(\beta,\gamma)\,\mathsf{Var}(\theta_{st}\,|s_{st}) - (1-\gamma)\,\mathsf{Var}(e_{lt}\,|s_{lt})\,. \end{aligned} \tag{20}$$

The last line in eq.(20) 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.

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.(7), we obtain that the first order conditions for the analyst's optimization problem at date 0 are (ignoring for the moment, the constraints that $0 \leq z_h \leq (\psi_h)^{-1} Z_h$ for $h \in \{st, lt\}$):

$$\begin{aligned} q(\beta,\gamma)\psi_{st} - 2az^*_{st} - cz^*_{lt} &= 0 \\ (1-\gamma)\psi_{lt} - 2bz^*_{lt} - cz^*_{st} &= 0 \end{aligned} \tag{21}$$

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.(10). 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 $0 \leq z_h \leq (\psi_h)^{-1} Z_h$ (with strict inequalities for an interior solution) and $4ab > c^2$.

The condition $z^*_h < (\psi_h)^{-1} Z_h$ is clearly always satisfied by setting Z_h 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:

$$\frac{\psi_{lt}}{\psi_{st}} \leq \frac{2b \times q(\beta,\gamma)}{c(1-\gamma)} \tag{22}$$

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

$$\frac{c \times q(\beta,\gamma)}{2a(1-\gamma)} \leq \frac{\psi_{lt}}{\psi_{st}}. \tag{23}$$

It is immediate that if Conditions (22) and (23) are satisfied then the condition $4ab > c^2$ is satisfied. Finally, it is easily checked that these two conditions are equivalent to:

$$c < \bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt}), \tag{24}$$

where:

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

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

$$\begin{aligned} \frac{\partial z^*_{st}}{\partial a} &= -\frac{4b}{(4ab-c^2)} z^*_{st} < 0, \\ \frac{\partial z^*_{lt}}{\partial a} &= \frac{2c}{(4ab-c^2)} z^*_{st} > 0 \quad \text{if} \quad c > 0. \end{aligned} \tag{25}$$

Proof of equations (12) and (13). 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: $\mathsf{Var}(\theta_{lt}\,|f^*_{lt}) = \mathsf{E}((\theta_{lt} - \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt}))^2 \mid f^*_{lt})$. The law of iterated expectations implies that $\mathsf{Var}(\theta_{lt}\,|f^*_{lt}) = \mathsf{E}(\mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})) \mid f^*_{lt})$. Since $\mathsf{Var}(\theta_{lt} \mid s_{st}, s_{lt})$ 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:

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

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 from eq.(7) that $\mathsf{Var}(\theta_{lt}\,|f^*_{lt}) = \beta^2(Z_{st} - \psi_{st}z^*_{st}) + (Z_{lt} - \psi_{lt}z^*_{lt})$. Therefore $\mathcal{I}_{lt}$ is as given in eq.(13). The derivation of the expression for $\mathcal{I}_{st}$ follows the same step and is omitted for brevity.

Proof of Corollary 1. Differentiating eq.(12) and eq.(13) with respect to the marginal

cost of producing short-term information, a, we obtain

$$\frac{\partial \mathcal{I}_{st}}{\partial a} = (\frac{\partial z^*_{st}}{\partial a})\psi_{st}\mathcal{I}^2_{st}, \tag{26}$$

and

$$\frac{\partial \mathcal{I}_{lt}}{\partial a} = (\beta^2\psi_{st}\frac{\partial z^*_{st}}{\partial a} + \psi_{lt}\frac{\partial z^*_{lt}}{\partial a})\mathcal{I}_{lt} = -(\frac{2(2\beta^2\psi_{st}b - c\psi_{lt})}{(4ab - c^2)})z^*_{st}\mathcal{I}^2_{lt}. \tag{27}$$

As $\frac{\partial z^*_{st}}{\partial a} < 0$ (see eq.(25)), eq.(26) implies that $\frac{\partial \mathcal{I}_{st}}{\partial a} < 0$. Moreover as $z^*_{st} > 0$, eq.(27) implies that $\frac{\partial \mathcal{I}_{lt}}{\partial a} > 0$ if and only if $\beta < (\frac{c\psi_{lt}}{2b\psi_{st}})^{\frac{1}{2}}$.

Online Appendix for

Does Alternative Data Improve Financial Forecasting? The Horizon Effect

(not intended for publication)

July 20, 2021

Contents

1 Dividing Forecasting Tasks

In our model, the analyst is in charge of two forecasting tasks and bears a multi-tasking cost. One may wonder whether she would not be better off assigning these two tasks to two different agents to save on the multitasking cost. In this section, we identify three reasons why this may not be optimal: (i) Duplication of fixed cost of information production, (ii) Agency costs and (iii) Imperfect communication between the agents. We provide conditions on the parameters in Section 1.1 and 1.3 such that dividing the tasks is suboptimal.

1.1 Duplication of fixed costs of information production

Suppose that the "analyst" assigns the tasks of forecasting short-term and long-term earnings to two different agents. We call these agents: (i) "*st*" (in charge of forecasting the short-term earnings) and (ii) "*lt*" (in charge of forecasting the long-term earnings). As in the baseline model, the *st*-agent obtains a signal $s_{st} = \theta_{st} + \varepsilon_{st}$ and can exert the effort z_{st} to reduce the variance of the noise in her signal and the $lt-$agent obtains a signal $s_{lt} = e_{lt} + \varepsilon_{lt}$ and can exert the effort z_{lt} to reduce the variance of the noise in her signal. The cost of effort for the *st*-agent is $C_{st}(z_{st}) = C_0 + a \times z_{st}^2$ and the cost of effort for the *lt*-agent is $C_{lt}(z_{lt}) = C_0 + b \times z_{lt}^2$ where C_0 is the fixed cost of acquiring information about the firm.

We first assume that the agents can truthfully, and costlessly, report their signals to the analyst (the principal). Moreover, there is no agency problem: agents' efforts are observable and the analyst perfectly controls the effort exerted by each agent. The compensation ω_j paid to the agent $j \in \{st, lt\}$ must be high enough to cover his effort cost. Thus, the participation constraint of agent $j \in \{st, lt\}$ is (his outside option is worth zero to simplify)

$$\omega_j \geq C_j(z_j),$$

and the analyst's final payoff (net of the compensation of the agents) is

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

Given the signals reported by the agents, the analyst forms her forecasts optimally, as in the

baseline model. Thus, proceeding as in the baseline model, the analyst's objective function at date 0 is to choose $\{z_{st}^{**}, z_{lt}^{**}, \omega_{st}^{*}, \omega_{lt}^{*}\}$ solving

$$\max_{\{z_{st}, z_{lt}, \omega_{st}, \omega_{lt}\}} \omega - \gamma Var(\theta_{st} \,|s_{st}) - (1-\gamma) Var(\theta_{st} \,|s_{st}, s_{lt}) - \omega_{st} - \omega_{lt}$$

$$u.c \;\; : \;\; \omega_j \geq C_j(z_j) \text{ for } j \in \{st, lt\},$$

Clearly, for fixed $\{z_{st}, z_{lt}\}$, it is optimal for the analyst to choose the lowest compensation for the agents, i.e., to set $\omega_j = C_j(z_j)$. We deduce that the analyst's objective function at date 0 is

$$\max_{\{z_{st}, z_{lt}\}} H(z_{st}, z_{lt}) = \omega - q(\beta, \gamma) Var(\theta_{st} \,|s_{st}) - (1-\gamma) Var(e_{lt} \,|s_{st}, s_{lt}) - 2C_0 - a \times z_{st}^2 - b \times z_{lt}^2. \quad (1)$$

There are two differences with the case considered in the baseline model. First, by assigning the forecasting tasks to two different agents, the analyst avoids the cost of multi-tasking, c. Second, the total fixed cost of acquiring information is $2C_0$ instead of C_0 because each agent must pay this cost.

Let $z_j^*(c)$ be the optimal effort when the cost of multi-tasking is c, as given in Proposition 1. Clearly, solving eq.(1) is identical to the analyst's problem in the baseline model when $c = 0$ (since C_0 does not depend on efforts). Thus, everything else being equal, we have: $z_j^{**} = z_j^*(0)$.[1] Note that $z_j^*(c) < z_j^*(0)$. Thus, the analyst requires higher efforts for each task from the agents because, with two agents, she saves on the cost of multi-tasking. As a result, the analyst's weighted forecasting error with two agents is smaller than in the baseline model.

However this does not mean that hiring two agents is optimal, because each agent must be compensated for the fixed cost of collecting information. In fact, the analyst is better off *not* dividing the tasks between two agents if (and only if):

$$J(z_{st}^*(c), z_{lt}^*(c)) \geq H(z_{st}^*(0), z_{lt}^*(0)), \quad (2)$$

[1] Observe that the condition on c in Proposition 1 is sufficient to guarantee that if the solution to the analyst's problem is interior when $c > 0$ then it is for $c = 0$.

where $J(z^*_{st}(c), z^*_{lt}(c))$ is defined in the text. Using the fact that $H(z^*_{st}(0), z^*_{lt}(0)) = J(z^*_{st}(0), z^*_{lt}(0)) + cz^*_{st}(0)z^*_{lt}(0) - C_0$, we can rewrite eq.(2) as:

$$C_0 - cz^*_{st}(0)z^*_{lt}(0) \geq J(z^*_{st}(0), z^*_{lt}(0)) - J(z^*_{st}(c), z^*_{lt}(c)). \tag{3}$$

The R.H.S is negative because $\{z^*_{st}(c), z^*_{lt}(c)\}$ maximizes J. Thus, a sufficient condition for Condition (2) to hold is that:

$$cz^*_{st}(0)z^*_{lt}(0) \leq C_0,$$

which, using the expressions for $z^*_{st}(0)$ and $z^*_{lt}(0)$ in Proposition 1, is equivalent to:

$$c \leq \frac{4C_0}{h(\beta,\gamma)(1-\gamma)\psi_{st}\psi_{lt}}. \tag{4}$$

Thus, we obtain that if $c \leq \text{Min}\{\bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt}), \frac{4C_0}{q(\beta,\gamma)(1-\gamma)\psi_{st}\psi_{lt}}\}$ (where $\bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt})$ is defined in the proof of Proposition 1), Proposition 1 holds and it is not optimal for the analyst to hire two agents, despite the multi-tasking cost.

1.2 Agency costs.

Agency frictions (e.g., if the analyst cannot perfectly observe the two agents' efforts) would add incentive compatibility constraints to the analyst's optimization problem. Hence, agency frictions can only reduce the maximum expected payoff for the analyst when she divides the task between two agents, $H(z^*_{st}(0), z^*_{lt}(0))$. Hence, Condition (4) is sufficient for the analyst being not better off dividing forecasting tasks between two agents when one introduces agency issues in the set-up considered in the previous section.

1.3 Complementarity and imperfect communication

The analysis in Section 1.1 implicitly assumes that the tasks of obtaining information about the common component and the unique component of the long-term earnings can be separated. A more plausible assumption is that achieving the first task is necessary to achieve the second one. Intuitively, one cannot obtain a signal about the unique component of the long-term earnings without first filtering out the common component from information

about the long-term earnings. The reverse is not true because information about the common component can be obtained by just focusing on information relevant for forecasting the short-term earnings. Thus, it is natural to see the tasks of obtaining signals about the common and the unique components in the firm's earnings as being "ordered." Achieving the first task (obtaining a signal about the short-term earnings, i.e., the common component of firms' earnings) is a necessary prerequisite for achieving the second one (obtaining a signal about the unique component of the long-term earnings). This ordering creates a form of complementarity between the two tasks: the second task yields a signal only if the first one has been completed.

To analyze this scenario, we consider a slightly different formulation of the information structure in our model. Suppose that the long-term signal is

$$\begin{aligned} \widehat{s}_{lt}(\iota) &= s_{st} + \eta + s_{lt} = e_{lt} + \theta_{st} + \varepsilon_{lt} + \varepsilon_{st} + \eta \quad \text{if} \quad \iota = 1, \\ \widehat{s}_{lt}(\iota) &= \emptyset \quad \text{if} \quad \iota = 0. \end{aligned} \tag{5}$$

where η has a normal distribution with mean zero and variance σ_η^2. The indicator variable ι is equal to 1 if the short-term signal s_{st} has been produced and 0 otherwise. This means that the long-term signal is observed if and only if the short-term signal is produced. If $\sigma_\eta^2 = 0$ and $\iota = 1$, this specification is equivalent to that considered in the model because, for the analyst, observing $\{s_{st}, \widehat{s}_{lt}\}$ is equivalent to observe $\{s_{st}, s_{lt}\}$. The case in which $\sigma_\eta^2 > 0$ can be interpreted as the case in which the short-term signal is observed with noise before producing the long-term signal.

With this specification for the short-term and the long-term signals, there are three possibilities to consider. The first possibility is the case in which the analyst does not divide the tasks, as in the baseline model. In this case, $\sigma_\eta^2 = 0$ if $\iota = 1$ because the analyst observes perfectly the short-term signal since she produces it. Choosing $\iota = 1$ is equivalent to choose to cover the firm and as in the baseline case, this is always optimal for ω large enough. Thus, we are back to the case analyzed in the paper in which the analyst's optimal expected payoff is $J(z_{st}^*(c), z_{lt}^*(c))$.

The second possibility is that the analyst hires the "st" and the "lt" agents. The first is

in charge of producing the signal s_{st} and the second is in charge of producing the long-term signal $\widehat{s}_{lt}(1)$. Both work independently and report their signals to the analyst. However, even if $\sigma_\eta^2 = 0$, this case cannot yield a higher expected payoff to the analyst than the previous case. Indeed, to produce the long-term signal, the long-term agent must first produce the short-term signal and pay the multitasking cost. Thus, the short-term signal is produced twice and the multitasking cost is paid anyway. It is therefore better for the analyst to directly produce the two signals to avoid duplication of efforts for the production of the short-term signal.

The third and most interesting possibility is the case in which the analyst delegates the forecasting tasks to two different agents and the two agents can communicate to avoid duplications of efforts in the production of the short-term signal. In this case, the *st*-agent first produces the short-term signal and then *communicates* this signal (s_{st}) to the *lt*-agent. If communication is perfect ($\sigma_\eta^2 = 0$), we are back to the case already analyzed in Section 1.1. However, a more realistic possibility is that communication between both agents is *imperfect* so that $\sigma_\eta^2 > 0$. In this case, the observation of $\{s_{st}, \widehat{s}_{lt}\}$ is equivalent to observing $\{s_{st}, s'_{lt}\}$ where $s'_{lt} = s_{lt} + \eta$. Thus, from the agent's reports, the analyst obtains a less precise long-term signal than when $\sigma_\eta^2 = 0$. Intuitively, the analyst cannot distinguish in the signal conveyed by the *lt*-agent what is due to noise arising from the lack of information about the unique component of the firm's long-term earnings (ε_{lt}) and what is due to noise in the communication between the agents (η).

If communication between the agents is costless, then $\iota = 1$ is optimal, i.e., it is optimal for the analyst to let the agents communicate even if communication is noisy (because without communication the *lt*-agent cannot obtain the long-term signal, unless he pays the cost of multi-tasking). In this case, the analyst's problem with two agents is given by eq.(1), replacing s_{lt} by s'_{lt} and

$$Var(e_{lt} \,|s_{st}, \widehat{s}_{lt}(1)) = Var(e_{lt} \,|s'_{lt}) = \sigma_\eta^2 + (Z - z_{st})\psi_{lt}.$$

The rest of the analysis is identical to that in Section 1.1 and after some algebra, we obtain

that if

$$c \leq \frac{4C_0 + (1-\gamma)\sigma_\eta^2}{q(\beta,\gamma)(1-\gamma)\psi_{st}\psi_{lt}}, \tag{6}$$

then the analyst is *better off not* splitting the production of the short-term and long-term signals between two agents. Note that this condition can be satisfied even if $C_0 = 0$, provided that the communication between the two agents is noisy ($\sigma_\eta^2 > 0$). The reason is that a single analyst better exploits the complementarity that naturally exists between the two tasks because there is no loss of information through communication (a single analyst perfectly communicates with herself).

2 Shock on ψ_{st}

In this Appendix, we show that if $\beta < \frac{1}{2}(\frac{c\psi_{lt}}{b\psi_{st}})^{\frac{1}{2}}$ then (i) the informativeness of the analyst's short-term forecast increases with the marginal return on effort for obtaining short-term information (ψ_{st}), i.e., $\frac{\partial \mathcal{I}_{st}}{\partial \psi_{st}} > 0$ and (ii) the informativeness of the analyst's long-term forecast decreases with the marginal return on effort for obtaining short-term information ($\frac{\partial \mathcal{I}_{lt}}{\partial \psi_{st}} < 0$).

First, it is direct from Proposition 1 that

$$\frac{\partial z_{st}^*}{\partial \psi_{st}} = \frac{2bq(\beta,\gamma)}{4ab - c^2} > 0, \quad \text{and} \quad \frac{\partial z_{lt}^*}{\partial \psi_{st}} = -\frac{cq(\beta,\gamma)}{4ab - c^2} < 0. \tag{7}$$

Thus, when ψ_{st} increases, the analyst exerts more effort to collect short-term information and less effort to collect long-term information. The mechanism is the same as for a decrease in the marginal cost of obtaining short-term information. Indeed, both types of shocks increase the marginal informational benefit of effort to collect short-term information. Thus, the analyst exerts more effort to collect short-term information (as the marginal benefit of effort decreases with effort). However, this raises the marginal cost of effort to collect long-term information when multi-tasking is costly ($c > 0$). Consequently, the marginal benefit of collecting long-term information declines. As the optimal allocation of effort requires equalizing the marginal benefit of effort on each task, the analyst optimally reacts by reducing her effort to collect long-term information.

Using eq.(12) in the main text, it immediately follows that $\frac{\partial \mathcal{I}_{st}}{\partial \psi_{st}} > 0$ because $\mathcal{I}_{st}$ increases in z^*_{st} and ψ_{st}. The effect of ψ_{st} on $\mathcal{I}_{lt}$ is negative if and only if the effect of ψ_{st} on $(\beta^2(Z_{st} - \psi_{st}z^*_{st})) + (Z_{lt} - \psi_{lt}z^*_{lt})$ is positive (see eq.(13) in the main text). A sufficient and necessary condition for this is that:

$$\beta^2(\psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}} + z^*_{st}) + \psi_{lt}\frac{\partial z^*_{lt}}{\partial \psi_{st}} < 0. \tag{8}$$

Substituting $\frac{\partial z^*_h}{\partial \psi_{st}}$ by its expression in eq.(7) and z^*_{st} by its expression in Proposition 1 in eq.(8), we deduce that a sufficient condition for this is $\beta < \frac{1}{2}(\frac{c\psi_{lt}}{b\psi_{st}})^{\frac{1}{2}}$.

3 EPS to Net Income Forecast Conversion

Converting an EPS forecast to a Net Income Forecast is not immediate because I/B/E/S does not report the number of shares used by the analyst to compute EPS. We experimented with two different approaches to make that conversion: (i) multiply the unadjusted EPS forecast from I/B/E/S by the number of shares from CRSP at t (*shrout*), or (ii) multiply the actual net income observed ex-post by the ratio of the I/B/E/S adjusted EPS forecast over the I/B/E/S adjusted actual EPS. This last 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, while preserving the ratio of forecast error relative to realized earnings reported in I/B/E/S.

To evaluate the quality of each approach, we compared the net income forecast obtained after converting the EPS forecast with the true net income forecast whenever the analyst issues both. For almost 60% of those cases, the difference (in absolute value) between the converted EPS and the true net income forecast is lower with the second approach, and so we retain this one for making this conversion whenever an EPS forecast is available but the net income forecast is not.

4 Why not Use Long-Term Growth Forecasts?

Analysts sometimes disclose, in addition to their earnings forecasts, a forecast about long-term growth. Specifically, a long-term growth (LTG) forecast in percent is reported instead of earnings forecasts in dollar amounts for more distant and specific fiscal periods. These LTG forecasts have been used in the literature, either directly to understand belief formation (e.g., Bordalo, Gennaioli, La Porta, and Shleifer (2019)), or indirectly to estimate the cost of capital (e.g., Chen, Da, and Xhao (2013)). LTG forecasts are however not well suited for our purpose because the horizon of these forecasts is unclear, making it difficult to assign them to actual realizations (and hence measure precisely their informativeness by horizon). After reading several reports from analysts, we indeed find substantial heterogeneity in how analysts define the horizon of their LTG forecasts (when they provide this definition). Some refer to earnings growth for the next five years, others use the next three years. Many refer to 3-5 year growth, without any further detail. Moreover, the base year for the growth estimate also varies. It can be the last historical fiscal year, the current fiscal year, the next fiscal year, or the subsequent one. Often, this base year is undefined.

5 Forecast Informativeness with Biased Analysts

To allow for the possibility of a systematic bias in the analyst's forecasts, suppose that these forecasts are given by:

$$f_{hi} = \mathrm{E}(\theta_{hi} \,|\Omega) + \widetilde{b}_{hi}, \tag{9}$$

where f_{hi} is the analyst's forecast about firm i's earnings, θ_{hi}, at horizon h, Ω is the analyst's information and $\widetilde{b}_{hi}$ is the analyst's bias, which can be random. In our model, $\widetilde{b}_{hi} = 0$ (the analyst is unbiased). On average, the analyst's bias at horizon h is

$$\mathrm{E}(\theta_{hi} - f_{hi}) = \mathrm{E}(\widetilde{b}_{hi}).$$

The literature on equity sell-side analysts suggests that $\mathrm{E}(\widetilde{b}_{hi}) \geq 0$. As explained in the text, the quality of analysts' forecasts is often measured by the average forecasting error. The

analyst's expected forecasting error is

$$\begin{aligned} \mathrm{E}((\theta_{hi} - f_{hi})^2) &= \mathrm{E}((\theta_{hi} - \mathrm{E}(\theta_{hi}\,|\Omega) + \mathrm{E}(\theta_{hi}\,|\Omega) - f_{hi})^2) \\ &= \mathrm{Var}(\theta_{hi}\,|\Omega) + \mathrm{E}(\widetilde{b}_{hi}^2) \\ &= \mathrm{Var}(\theta_{hi}\,|\Omega) + \mathrm{Var}(\widetilde{b}_{hi}) + \mathrm{E}(\widetilde{b}_{hi})^2 \end{aligned}$$

Thus, when an analyst is biased, her expected forecasting error is the sum of: (i) the precision of her forecast ($\mathrm{Var}(\theta_{hi}\,|\Omega)$), (ii) the variance of her bias, or (iii) her expected bias ($\mathrm{E}(\widetilde{b}_{hi})$). In contrast, as shown below, our measure of the analyst's forecast informativeness is not affected by the expected bias and identical to the informativeness of the analyst's unbiased forecast when $\mathrm{Var}(\theta_{hi}\,|\Omega) = 0$.

To see this, let denote by f^*_{hi} the analyst's unbiased expected forecast: $f^*_{hi} = \mathrm{E}(\theta_{hi}\,|\Omega)$. Assuming that all variables have a normal distribution, we have

$$\mathrm{E}(\theta_{hi}\,|f_{hi}) = \widehat{k}_0 + \widehat{k}_1 f_{hi}$$

with $\widehat{k}_0 = (E(\theta_{hi})(1 - \widehat{k}_1) - \widehat{k}_1 \mathrm{E}(\widetilde{b}_{hi}))$ and $\widehat{k}_1 = \frac{Var(f^*_{hi})}{Var(f_{hi})}$. Assuming, as we do in our tests, that the observations of (θ_{hi}, f_{hi}) for different firms are independent draws from the same distribution, the estimate of k_1 (k_0) in the regression considered in eq.(14) in our paper is a (consistent) estimate of $\widehat{k}_1$ ($\widehat{k}_0$). The R^2 of this regression is our measure of an analyst's forecast informativeness at horizon h. Its theoretical value is

$$R^2_{ih} = \widehat{k}_1^2 \frac{Var(f_{hi})}{Var(\theta_{hi})} = \widehat{k}_1 R^2_{\theta f^*} \tag{10}$$

where $R^2_{\theta f^*}$ is the R^2 of a regression of θ_{hi} on f^*_{hi}. Thus, our measure of informativeness is not affected by the expected level of the bias in the analyst's forecast ($\mathrm{E}(\widetilde{b}_{hi})$) in contrast to the expected forecasting error. Moreover, if the analyst's bias is constant across firms ($\mathrm{Var}(\widetilde{b}_{hi}) = 0$), our empirical measure of the informativeness of an analyst's forecast is identical to the the informativeness of the analyst's *unbiased forecast*, f^* (which is not observed). Indeed, in this case, $\hat{k}_1 = 1$ so that $R^2_{ih} = R^2_{\theta f^*}$. If instead $\mathrm{Var}(\widetilde{b}_{hi}) > 0$, our empirical measure is biased downward (it underestimates the true informativeness of analysts' unbiased forecasts

at a given horizon). However, there is a one-to-one mapping between our empirical measure of forecast informativeness (R^2_{ih}) and the informativeness of the analyst's unbiased forecast ($R^2_{\theta f*}$).

6 Analysts' Forecasting Activity and Recommendations

Our second test (Test#2) builds on the assumption that (some) analysts use StockTwits data as a complementary source of information (see discussion in Section VI.B). Table A1 and Table A2 report results (discussed in section VI.B) that are consistent with this assumption.

In Table A1, Column (1) shows that analysts are more likely to issue a new forecast on a given firm and day following an increase in StockTwits activity, as measured by the number of actual messages posted about the firm over the last 30 days. Column (2) shows that this result survives when controlling for trading volume, and thus for the possible effects of contemporaneous news (public or private) that is material enough to generate trading. Columns (3) and (4) show that this result continues to hold on days without news arrival from traditional data sources (which we identify using Capital IQ Key Developments), and thus mitigate the concern that news arrival (affecting both analysts' forecasts and social media activity) confounds the relationship documented in Column (1).

In Table A2, Column (1) shows that the recommendation of an analyst on a given firm and day is positively (negatively) related to the fraction of StockTwits users whose opinion is "Bullish" ("Bearish"). When more users are "Bullish" ("Bearish"), analysts are more likely to upgrade (downgrade) their recommendation. The economic magnitude of this effect is small, but it is highly significant. Columns (2) and (3) show that this result continues to hold on days without news arrival from traditional data sources (which we identify using Capital IQ Key Developments), and thus mitigate the concern that news arrival (affecting both analysts' recommendations and users' ratings) confounds the relationship documented in Column (1) of Table A2.

Table A1: Social Media Data and Analysts' Forecasting Activity

This table presents OLS estimates of the sensitivity of analysts' propensity to issue new earnings forecasts to recent StockTwits activity. Estimations are made at the analyst-firm-day level. The sample includes all U.S. 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 otherwise. #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 StockTwits. Trading Volume is the total volume of trading on from $t-30$ to $t-1$. In Column (3), we impose that no news (from the Capital IQ Key Developments 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). 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 without news in Key Dev. at t	No	No	Yes	No
Sample without news in Key Dev. over $_{t-30\to t}$	No	No	No	Yes
N	80,434,931	80,379,362	69,414,958	3,147,979

Table A2: Social Media Data and Analysts' Recommendations

This table presents OLS estimates of the sensitivity of analysts' recommendations to the number of "Bullish" and "Bearish" ratings issued by StockTwits users. Estimations are made at the analyst-firm-day level. The sample includes all U.S. firms covered by at least one analyst between 2009 and 2017. The dependent variable is the last available recommendation made by analyst i on firm j at t (measured by the item ireccd in I/B/E/S and multiplied by -1 so that greater values of ireccd indicate better recommendations). *Rating* is the difference between the fraction of "Bullish" users and that of "Bearish" users about j at $t-1$. *Rating* is naturally bounded between -1 (all users are "Bearish") and +1 (all users are "Bullish"). We require that there are at least 10 users with an active rating about j. A rating is active if it is the last available rating, and if it is not stale at $t-1$. A rating is stale after 365 days. In Column (2), we impose that no news (from the Capital IQ Key Developments dataset) is released about the firm during the day (otherwise the observation is removed from the sample), i.e., at t. In Column (3), we impose that no news is released about the firm from $t-30$ to t (otherwise the observation is removed from the sample). 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:	Analyst Recommendation		
OLS:	(1)	(2)	(3)
Rating	0.11***	0.11***	0.14***
	(6.89)	(7.13)	(4.22)
Analyst × Firm FE	Yes	Yes	Yes
Analyst × Date FE	Yes	Yes	Yes
Sample without news in Key Dev. at t	No	Yes	No
Sample without news in Key Dev. over $_{t-30\to t}$	No	No	Yes
N	33,758,191	28,677,022	879,011

7 Actual Messages vs. Hypothetical Messages

This appendix compares actual and hypothetical messages, and decomposes the sources of variation for each variable. The number of actual messages (*#Messages*) about firm j on day t (after coverage initiation by StockTwits) can be decomposed as

$$\begin{aligned} \#Messages_{j,t} &= \frac{\#Messages_{j,t}}{\#Total\ Messages_t} \times \#Total\ Messages_t \\ &= w_{j,t} \times \#Total\ Messages_t \end{aligned}$$

where $w_{j,t}$ is the share (in percentage) of total messages posted on StockTwits about j at t, and $\#Total\ Messages_t$ is the total number of messages posted on the platform on day t. Assuming (for convenience) that analyst i covers only firm j, the actual messages she is exposed to, denoted $\#Messages_{i,t}$, can be decomposed as:

$$\#Messages_{i,t} = w_{i,t} \times \#Total\ Messages_t \times Post_{i,t}, \tag{11}$$

where $w_{i,t} = w_{j,t}$ (because i only follows j), and $Post_{i,t}$ is an indicator equal to one after firm j is discussed on StockTwits for the first time ($\#Messages_{i,t}$ is set to zero before coverage by StockTwits begins). Variation in analyst i's exposure ($\#Messages_{i,t}$) is the product of three components: (i) the relative cross-sectional variation in the share of messages analyst i is exposed to, captured by $w_{i,t}$, (ii) the aggregate variation of total messaging on StockTwits captured by $\#Total\ Messages_t$, and (iii) time variation due to the staggered introduction of StockTwits, captured by $Post_{i,t}$.

Using a similar decomposition, exposure based on hypothetical messages is given by the following product:

$$\#Hypothetical\ Messages_{i,t} = w'_i \times \#Total\ Messages_t \times Post_{i,t}, \tag{12}$$

where $w'_i = \overline{w_j}$ is the average of $w_{j,t}$ across all t, after messaging about firm j begins.[2]

[2]Using other methodologies to estimate hypothetical messages does not materially affect our results. For example, one could use the median (rather than the average) of $w_{j,t}$ to compute w'_i, or use $Post'_t$ instead of $Post_{i,t}$, where $Post'_t$ is equal to one after January 1, 2009.

Comparing eq.(12) with eq.(11) highlights that the first component (i.e., w'_i) in eq.(12) is time-invariant. Thus, while exposure based on $\#Messages_{i,t}$ could capture variation unrelated to StockTwits (e.g., if the arrival of information about firm j from other sources than StockTwits at t correlates with $w_{i,t}$ $(= w_{j,t})$ because StockTwits' users relay or comment that information), $\#Hypothetical\ Messages_{i,t}$ cannot because the share w'_i is fixed (and thus cannot vary with such information arrival). Of course, $\#Hypothetical\ Messages_{i,t}$ still captures variation across firms via w'_i (i.e., some analysts follow firms that are systematically more discussed), but this variation is controlled for by the analyst fixed effects η_i in our tests. Therefore, the source of variation we use in the paper to estimate the effect of greater exposure to StockTwits' data based on hypothetical messages comes solely from heterogeneous exposure to the progressive and staggered expansion of the platform (measured by $\#Total\ Messages_t \times Post_{i,t}$).[3]

Although our presentation focuses on the case where analyst i follows only firm j, the source of variation that our test relies upon is the same when analysts cover several firms, if coverage is stable. Since coverage is persistent on average, most changes in w'_i (i.e., the average of $\overline{w_j}$ across the covered firms j) will be captured by the fixed effects η_i, and the main source of variation will come from the *aggregate* variation in the number of messages (and from the staggered deployment of the platform). To mitigate the concern that changes in analyst coverage (i.e., change in w'_i over time) could explain our results, we verify and show that our estimates are not materially affected when we focus on the sub-sample of analysts covering always the same firms (see Table A9 in Section 11 of this Appendix). Alternatively, the variation in w'_i that is not fully captured by η_i due to changes in coverage can be directly controlled for in the regression. The share $\overline{w_j}$ is indeed perfectly observed for all firms because we use it to compute hypothetical messages. We average this variable across firms by analyst, day, and horizon to obtain w'_i. Table A3 shows that controlling for w'_i leads to similar conclusions.

[3]Put it differently, $\#Hypothetical\ Messages_{i,t}$ captures three sources of variation related to treatment: (i) w'_i, measuring the degree of exposure to treatment, (ii) $\#Total\ Messages_t$, measuring the overall treatment intensity, and (iii) $Post_{i,t}$, measuring the treatment status. This third and last source of variation is the same as the one used to identify treatment in a standard staggered diff-in-diff specification. Since the first source of variation is absorbed by η_i in eq.(18), only the last two contribute to the estimation.

Table A3: Controlling for Analysts' Average Share of All Messages (w_i')

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on 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 a variable capturing the 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 forecast 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 hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). w_i' is the mean of $\overline{w_j}$ across the firms covered by the analyst. $\overline{w_j}$ is the mean of $w_{j,t}$ for all t after a message is observed for the first time about j. 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 $t-1$. In columns (2), (3), analyst and date fixed effects are interacted with h^*. Detailed variable definitions are provided in Appendix II. 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 Proxy:	# Hypothetical Messages		
OLS:	(1)	(2)	(3)
$h^*\times$ Data Exposure	-1.15***	-1.03***	-1.05***
	(-3.72)	(-4.40)	(-5.03)
Data Exposure	0.05	-0.39	-0.4
	(0.28)	(-1.57)	(-1.60)
$h^*\times w_i'$	3.54***	0.33	-0.21
	(2.37)	(0.20)	(-0.13)
w_i'	2.79***	3.46***	1.3
	(2.42)	(2.62)	(0.93)
h^*	-16.77***		
	(-32.69)		
Analyst FE	Yes		
Date FE	Yes		
Analyst FE (interacted)		Yes	Yes
Date FE (interacted)		Yes	Yes
Controls			Yes
N	30,959,276	30,105,551	27,860,424

8 Do Our Measures Correlate with News from Standard Sources?

Our second test (Test#2) builds on the assumption that our two measures of analysts' exposure to StockTwits' data ("Data Exposure") do not correlate with the regular flow of firm-level information coming from standard sources (see discussion in Section VI.C). Tables A4 and A5 present the results of two tests (mentioned in section VI.C) attempting to falsify this assumption.

We use Capital IQ Key Developments to identify the regular flow of firm-level information from standard sources. This database is well-suited for two reasons. First, it covers a large spectrum of news category (e.g., announcements of earnings, dividend, M&As, executive changes, or SEC inquiries). There are almost 12 million news items in Capital IQ Key Developments about firms in our sample.[4] Second, the vast majority of the reported news items originate from standard sources (e.g., press releases, news wires, regulatory filings), which is precisely the news we want to identify (i.e., coming from "traditional" data). We use two approaches to measure the regular flow of firm-level information. First, we simply count the number of news items reported in Capital IQ about a given firm and time period (henceforth the "Volume Approach"). Second, we calculate the market response to each news item in absolute value, and use the sum for a given firm and time period to capture the relevance of these news items (henceforth the "Market Response Approach").[5] We then test whether these two measures of the flow of information for a given firm correlates with our measures of "Data Exposure".

Table A4 shows the results based on the "Volume Approach". We find no significant relationship between the *number* of daily news items reported in Capital IQ and the number of (i) users in a firm's watchlist (Columns (1) to (3)), or (ii) hypothetical messages (Columns (4) to (6)). As expected, however, we find a positive correlation with the number of actual messages (Columns (7) to (9)). Our assumption is thus rejected for this variable, but it

[4] In our tests, we consider all news except M&A rumors, because these rumors may actually come from social media outlets.

[5] We set this sum to zero when no news is reported.

is *not* rejected for the two measures of data exposure we use. Table A5 shows similar results based on the "Market Response Approach" instead of the number of news. In sum, neither the number of news items arriving from standard sources, nor their relevance correlate significantly with either a firm's watchlist, or hypothetical messages.

Table A4: Data Exposure and News Arrival (Volume Approach)

This table presents OLS estimates of the sensitivity of different measures of social media data exposure to news arrival from standard sources. Estimations are made at the firm-day level. 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 (3), *#Watchlist* is the number of StockTwits users having the firm in their watchlist on day t. In columns (4) to (6), *#Hypothetical Messages* is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (7) to (9), *#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 Developments on day $_t$. $\#News_{t\to T}$ is the number of distinct news about the firm reported in Capital IQ Key Developments between day t and day T. Capital IQ Key Developments 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*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\#\text{News}_t$	-4.66			-2.82			5.67***		
	(-0.59)			(-0.82)			(2.97)		
$\#\text{News}_{t-1}$		-3.98			-1.95			9.11***	
		(-0.51)			(-0.59)			(4.65)	
$\#\text{News}_{t-30\to t-1}$			-2.73			-1.68			10.06***
			(-0.41)			(-0.61)			(5.91)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	18,664,998	18,661,528	18,560,734	18,664,998	18,661,528	18,560,734	18,664,998	18,661,528	18,560,734

Table A5: Data Exposure and News Arrival (Market Response Approach)

This table presents OLS estimates of the sensitivity of different measures of social media data exposure to news arrival from standard sources. Estimations are made at the firm-day level. 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 (3), *#Watchlist* is the number of StockTwits users having the firm in their watchlist on day t. In columns (4) to (6), *#Hypothetical Messages* is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (7) to (9), *#Messages* is the number of actual messages posted about the firm from $t-30$ to $t-1$. Market Response to $\#\text{News}_t$ is the Absolute (value of the) Cumulative Abnormal Return ($ACAR_{j,t}$) observed in response to news about firm j reported in Capital IQ Key Developments on day t. Market Response to $\#News_t$ is set to zero when no news is reported. The cumulative abnormal return at t is computed with a two-day window $[t+0, t+1]$, using CRSP value-weighted index as a benchmark. Market Response to $\#News_{t\to T}$ is sum of all $ACAR_{j,t}$ observed in response to each news event about j reported in Capital IQ Key Developments between day t and day T. This variable is set to zero when no news is reported between t and T. Capital IQ Key Developments 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*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mkt Resp. to $\#\text{News}_t$	1.35			-0.44			5.14***		
	(0.80)			(-0.44)			(6.38)		
Mkt Resp. to $\#\text{News}_{t-1}$		1.60			-0.32			6.56***	
		(0.95)			(-0.33)			(7.74)	
Mkt Resp. to $\#\text{News}_{t-30\to t-1}$			-0.30			-0.67			4.91***
			(-0.26)			(-0.98)			(10.31)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902

9 Robustness Table II

This Appendix discusses the robustness of the results reported in Table II (Section V.A). All robustness tests are reported in Table A6.

First, we find similar results in Panels A, B, and C when adding controls for various characteristics of the portfolio covered by the analyst. In Panel A, we report specifications that include fixed effects for two-digit SIC industries.[6] In Panel B, we further control for the average characteristics of the covered firms, namely: size (log of total assets), (log of) age, cash flow to assets, debt to assets, cash to assets, and Tobin's Q. Finally, Panel C shows similar results using the same specification, but after we re-compute R^2 focusing only on forecasts about S&P500 firms, whose underlying characteristics have remained stable over time (Bai et al. (2016)).

Second, we show that the results are robust to focusing on analysts (Panel D) and firms (Panel E) for which both short and long-term forecasts are available. In Panel D we restrict the analysis to analysts who have issued at least one forecast with horizon greater than 3 years. In Panel E, we re-compute the dependent variable R^2 using only forecasts about firms for which at least one forecast with horizon greater than 3 year is available.

Finally, we check that our results are not specific to using the period 1983-1992 as our baseline, nor driven by I/B/E/S imperfect coverage at the beginning of the sample (Panel F). We also show that neither the number of forecasts used to estimate R^2 (Panel G), nor the assumptions we make about the updating speed of those forecasts (Panel H), materially affects inferences. Panel G reports specifications that include fixed effects for the number of observations used to estimate R^2 in eq.(14). Panel H reports results after we re-compute R^2 assuming analysts constantly update their forecasts. Specifically, we estimate an updated forecast every day, unless the analyst discloses one. We do so by linear interpolation between two consecutive disclosures for each analyst, firm, and fiscal period. This alternative approach for computing R^2 relaxes the implicit assumption that analysts update their forecasts only when we observe a new forecast.

[6] The constant is omitted because it is absorbed by the fixed effects.

Table A6: Robustness: Forecast Informativeness 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. 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 cumulative increment in R^2 over the 1993-2017 period. Variable definitions are in Appendix II. 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:	$0 < h \leq 1$	$1 < h \leq 2$	$2 < h \leq 3$	$3 < h \leq 4$	$4 < h \leq 5$
OLS:	(1)	(2)	(3)	(4)	(5)
Panel A: Controlling for changes in industry composition					
Year Trend	12.2***	11.1***	2.0	-7.6***	-14.2***
	(8.97)	(7.75)	(1.21)	(-3.15)	(-3.52)
Industry FE	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No
N	33,386,528	25,044,127	5,359,098	1,349,651	703,653
Panel B: Controlling for the characteristics of covered firms					
Year Trend	10.9***	8.5***	1.9	-5.0*	-9.2**
	(7.72)	(6.13)	(1.09)	(-1.70)	(-2.01)
Industry FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
N	31,175,295	23,216,441	4,994,926	1,286,975	670,362
Panel C: Focusing on SP500 firms					
Year Trend	11.8***	11.4***	5.9***	-4.2	-9.1**
	(6.35)	(5.50)	(2.61)	(-1.51)	(-2.03)
Constant (83-92)	80.1***	64.3***	56.3***	53.5***	49.6***
	(64.88)	(56.49)	(46.87)	(38.73)	(25.43)
N	18,423,237	14,206,102	3,138,963	769,951	406,058
Panel D: Analysts with both short and long-term forecasts					
Year Trend	6.9***	6.1***	1.6	-11.5***	-20.0***
	(4.78)	(4.14)	(0.85)	(-5.12)	(-5.41)
Constant (83-92)	78.6***	58.3***	47.4***	44.3***	42.6***
	(84.31)	(60.25)	(32.74)	(29.78)	(21.12)
N	8,600,935	7,389,585	3,663,585	1,349,749	703,712

Table A6: Robustness: Forecast Informativeness by Horizon (Cont'd)

Dep. variable:	Forecast informativeness (R^2)				
Sample:	$0 < h \leq 1$	$1 < h \leq 2$	$2 < h \leq 3$	$3 < h \leq 4$	$4 < h \leq 5$
OLS:	(1)	(2)	(3)	(4)	(5)
Panel E: Firms with both short and long-term forecasts					
Year Trend	7.6***	3.6**	0.1	-11.5***	-20.1***
	(4.86)	(2.23)	(0.08)	(-5.04)	(-5.38)
Constant (83-92)	78.6***	60.8***	50.2***	44.5***	42.7***
	(79.65)	(69.16)	(41.17)	(29.58)	(20.98)
N	29,023,675	22,491,017	5,159,145	1,338,504	698,958
Panel F: Excluding 80's					
Year Trend	7.6***	8.5***	3.5*	-11.8***	-18.3***
	(6.19)	(5.54)	(1.72)	(-4.66)	(-4.78)
Constant (90-92)	77.4***	55.6***	47.1***	44.5***	41.4***
	(113.19)	(62.63)	(31.24)	(26.28)	(19.77)
N	29,047,461	22,334,402	5,169,002	1,308,876	683,413
Panel G: Controlling for the number of observations used to compute R^2					
Year Trend	12.0***	10.2***	6.4***	-11.5***	-18.3***
	(8.33)	(7.25)	(3.46)	(-5.12)	(-5.22)
#Firms FE	Yes	Yes	Yes	Yes	Yes
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712
Panel H: Using R^2 based on interpolated forecasts					
Year Trend	9.8***	6.9***	-1.4	-11.1***	-13.4***
	(6.84)	(5.28)	(-1.30)	(-5.32)	(-3.98)
Constant (83-92)	78.2***	61.0***	56.1***	53.5***	50.9***
	(97.82)	(102.57)	(69.25)	(39.51)	(25.75)
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712

10 Robustness Table III

This Appendix discusses the robustness of the results reported in Table III (Section V.B). All robustness tests are reported in Table A7.

In Panel A, we report specifications controlling for the average characteristics of the covered firms, namely: size (log of total assets), (log of) age, cash flow to assets, debt to assets, cash to assets, and Tobin's Q.[7] Specifically, we average those average characteristics by (two-digit SIC) industry and year in Columns (2) and (3), and by analyst and year in Columns (4) and (5), and control for those in the regression.

Next, we verify that the results are also robust to focusing on analysts (Panel B) and firms (Panel C) for which both short and long-term forecasts are available. In Panel B we restrict the analysis to analysts who have issued at least one forecast with horizon greater than 3 years. In Panel C, we re-compute the dependent variable R^2 using only forecasts about firms for which at least one forecast with horizon greater than 3 year is available.

Finally, we show that neither the choice of our baseline period (Panel D), nor the assumptions we make about the updating speed of analysts forecasts (Panel E), materially affects our conclusions. In Panel D, we exclude the 80's and use the period 1990-1992 as our baseline. In Panel H, we re-compute R^2 assuming analysts constantly update their forecasts. Specifically, we estimate an updated forecast every day, unless the analyst discloses one. We do so by linear interpolation between two consecutive disclosures for each analyst, firm, and fiscal period. This alternative approach for computing R^2 relaxes the implicit assumption that analysts update their forecasts only when a new forecast is publicly disclosed.

[7] We do so in Columns (2) to (5), but not in Column (1) because we have too few observations of yearly slope estimates.

Table A7: The Slope of the Term Structure

This table presents OLS estimates of time trend in the slope of the term structure of forecasts' informativeness. The dependent variable is the slope of the term structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. A negative slope indicates that forecasts' informativeness (R^2) 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 cumulative change in slope over the 1993-2017 period. Variable definitions are in the Appendix II. 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.

Dep. variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS:	(1)	(2)	(3)	(4)	(5)
Panel A: Controlling for covered firms characteristics					
Year Trend	-10.8***	-4.7***	-4.4***	-4.2***	-2.8**
	(-6.74)	(-3.96)	(-3.28)	(-6.05)	(-2.22)
Constant (83-92)	-6.5***	-18.4***		-18.6***	
	(-6.45)	(-5.44)		(-7.70)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
Controls	-	Yes	Yes	Yes	Yes
N	33	1,083	1,080	7,256	6,909
Panel B: Focusing on SP500 firms					
Year Trend	-7.5***	-1.4	-2.5*	-5.2***	-3.4**
	(-3.60)	(-1.13)	(-1.80)	(-7.94)	(-2.07)
Constant (90-92)	-7.5***	-11.5***		-9.9***	
	(-5.67)	(-16.23)		(-22.22)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	803	772	4,533	4,307
Panel C: Analysts with short and long-term forecasts					
Year Trend	-10.1***	-4.5***	-2.8**	-4.9***	-2.7**
	(-6.20)	(-3.63)	(-2.30)	(-7.60)	(-2.07)
Constant (83-92)	-7.3***	-11.7***		-12.1***	
	(-7.06)	(-21.91)		(-25.95)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	7,657	7,290

Table A7: The Slope of the Term Structure (Cont'd)

Dep. variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS:	(1)	(2)	(3)	(4)	(5)
Panel D: Firms with short and long-term forecasts					
Year Trend	-9.4***	-3.7***	-2.6**	-4.4***	-2.5*
	(-5.62)	(-3.12)	(-2.12)	(-6.58)	(-2.83)
Constant (83-92)	-7.8***	-12.3***		-12.5***	
	(-7.41)	(-17.74)		(-25.19)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,050	1,019	7,619	7,252
Panel E: Excluding 80's					
Year Trend	-7.6***	-3.8***	-2.4**	-4.1***	-2.6*
	(-7.20)	(-3.90)	(-2.38)	(-5.20)	(-1.96)
Constant (90-92)	-8.5***	-12.0***		-12.7***	
	(-12.60)	(-23.33)		(-22.23)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	26	959	957	7,430	7,054
Panel F: Using R^2 based on interpolated forecasts					
Year Trend	-8.7***	-4.1***	-3.4***	-5.6***	-4.1***
	(-6.17)	(-4.60)	(-3.59)	(-8.07)	(-3.04)
Constant (83-92)	-5.3***	-9.3***		-8.9***	
	(-6.30)	(-21.47)		(-21.42)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	7,657	7,290

11 Robustness Table VI

This Appendix discusses the robustness of the results reported in Table VI (Section VI.D). Table A8 shows that our results are robust to controlling for trading volume and thus for the effects of news (public and private) that are material enough for generating trading. Table A9 shows that our results are also robust to focusing on analysts with stable coverage, and thus that changes in coverage cannot be the main explanation for our findings. Finally, we verify that focusing on analysts (Table A10) and firms (Table A11) for which both short and long-term forecasts are available does not affect inferences. Table A10 repeats the analysis focusing on analysts who have issued at least one forecast with horizon greater than 3 years. Table A11 does the same, but after we re-calculate R^2 using only forecasts about firms for which at least one forecast with horizon greater than 3 years is available.

Table A8: Robustness: Controlling for Trading Volume

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h-1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). 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 $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. Detailed variable definitions are provided in Appendix II. 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)
$h^*\times$ Data Exposure	-1.09***	-0.86***	-1.00***	-1.01***	-1.06***	-1.13***
	(-3.23)	(-3.17)	(-3.74)	(-3.88)	(-4.84)	(-5.32)
Data Exposure	0.16	-0.17	-0.3	0.38*	-0.14	-0.25
	(0.66)	(-0.68)	(-1.15)	(1.62)	(-0.62)	(-1.03)
$h^*\times$ Trading Volume	1.13***	0.62***	0.57***	1.18***	0.71***	0.66***
	(6.56)	(3.28)	(2.67)	(6.82)	(3.76)	(3.17)
Trading Volume	-0.4	-0.12	-1.23***	-0.43	-0.12	-1.23***
	(-1.29)	(-0.49)	(-3.80)	(-1.39)	(-0.48)	(-3.83)
h^*	-17.62***			-17.59***		
	(-31.69)			(-30.94)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,276	30,105,551	27,860,424	30,959,276	30,105,551	27,860,424

Table A9: Robustness: Analysts With Stable Coverage

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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)
$h^*\times$ Data Exposure	-0.46	-0.50**	-0.69***	-0.29	-0.71***	-0.85***
	(-1.49)	(-2.02)	(-2.60)	(-1.26)	(-3.46)	(-3.82)
Data Exposure	0.32	0.04	-0.15	0.48*	0.00	-0.16
	(1.25)	(0.15)	(-0.52)	(1.68)	(0.01)	(-0.64)
h^*	-16.35***			-16.34***		
	(-36.86)			(-35.24)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	14,552,288	13,773,488	12,683,367	14,552,288	13,773,488	12,683,367

Table A10: Robustness: Analysts With Non-Missing Long-Term Forecasts

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with non-missing long-term forecasts. An analyst has non-missing long-term forecasts if there is at least one non-missing $R^2_{i,t,h}$ for $h \geq 3$ over the sample period (2005-2017). 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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)
$h^*\times$ Data Exposure	-1.40***	-1.07***	-1.25***	-1.10***	-1.19***	-1.27***
	(-4.17)	(-3.34)	(-4.13)	(-4.59)	(-5.96)	(-7.42)
Data Exposure	-0.12	-0.29	-0.48	0.16	-0.26	-0.39
	(-0.54)	(-1.00)	(-1.49)	(0.71)	(-1.08)	(-1.58)
h^*	-15.33***			-15.24***		
	(-41.16)			(-38.40)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	13,782,999	13,019,477	12,153,633	13,782,999	13,019,477	12,153,633

Table A11: Robustness: Firms With Non-Missing Long-Term Forecasts

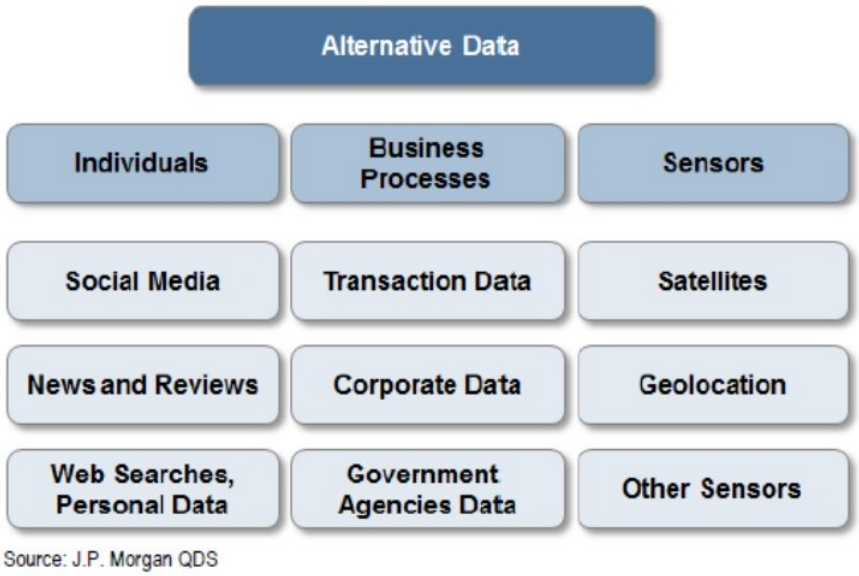
This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts covering firms with non-missing long-term forecasts. A firm has non-missing long-term forecasts if it has at least one non-missing forecast for $h \geq 3$ over the sample period (2005-2017). 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 a variable capturing the 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 forecast 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 average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h-1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. 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 Appendix II. 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)
$h^* \times$ Data Exposure	-0.86***	-0.78***	-0.96***	-0.69***	-0.94***	-1.05***
	(-2.59)	(-3.06)	(-3.72)	(-2.75)	(-4.54)	(-5.03)
Data Exposure	0.13	-0.17	-0.35	0.34	-0.14	-0.32
	(0.50)	(-0.64)	(-1.29)	(1.42)	(-0.57)	(-1.30)
h^*	-16.66***			-16.62***		
	(-33.85)			(-32.13)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,281	30,105,556	27,860,429	30,959,281	30,105,556	27,860,429

12 Alternative Data: Definition and Classification

Alternative data refers to any data containing relevant information about the value of firms that is not directly disclosed by them. These data sources can be broadly classified into three categories depending on whether they are produced by individuals (e.g. social media posts), generated through business processes / new technologies (e.g., credit card data or app data), or produced by sensors (e.g., satellite). This classification follows that of J.P.Morgan (Source: 2019 Handbook of Alternative Data, J.P.Morgan (Oct. 25, 2019)). It is summarized in their Figure 1 ("Classification of big/alternative data sources") on page 6, which we reproduce below.

Figure 1: Classification of big/alternative data sources



Source: J.P. Morgan QDS

Data generated by individuals include data from social media (e.g., Twitter, StockTwits, Facebook), from business-reviewing websites (e.g., Yelp) and E-commerce groups (e.g., Amazon), as well as web searches data (e.g., Google Search trends). Most of these data come in a text format. Data generated by business processes / new technologies include credit card data, supermarket scanner data, supply chain data, and app data, among others. Data generated by sensors typically include satellite imagines and geolocation data in general, as well as weather, natural disasters and pollution data.

13 Example of Analysts Using Social Media Data

J.P.Morgan

Completed 05 Dec 2017 12:26 AM EST
Disseminated 05 Dec 2017 12:27 AM EST
North America Equity Research
05 December 2017

AWS Partners Meeting Key Takeaways

Partner Lunch Confirms Our Positive View On re:Invent Announcements; Fargate & VMW Partnership Highlighted

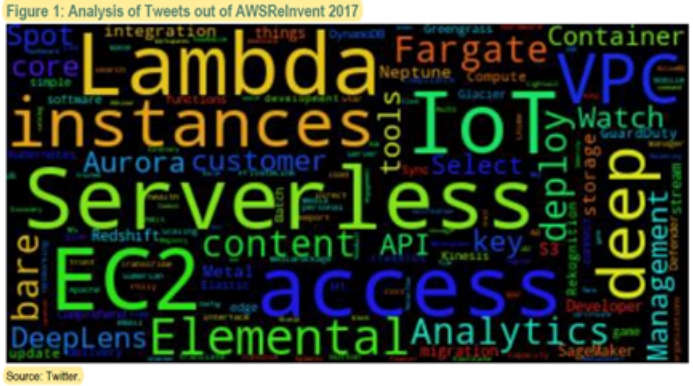
Internet
Doug Anmuth AC
(1-212) 622-6571
douglas.anmuth@jpmorgan.com
Bloomberg JPMA ANMUTH <GO>
Ashwin Kesireddy
(1-415) 315-6756
ashwin.x.kesireddy@jpmorgan.com
Cory A Carpenter
(1-212) 270-8125
cory.carpenter@jpmorgan.com
Software
Mark R Murphy AC

Coming out of the AWS re:Invent conference last week in Las Vegas, we feel more confident about Amazon's approach and ability to grow the AWS customer base. **We continue to believe AWS maintains a strong leadership position with an estimated ~75% market share, though we acknowledge MSFT Azure is gaining traction especially with larger enterprises and we estimate has a 15-20% market share.** In this note, we provide key takeaways from our lunch meeting with partners in the cloud ecosystem (hosted on Thursday 11/30) and other events we attended at the conference as a follow up to our Day 1 recap note. In addition, we analyzed ~22k tweets coming out of AWS re:Invent and found that Lambda/Serverless, EC2, IoT, VPC, Fargate, Deep Learning, and DeepLens were some of the top mentioned topics at the event.

- **Amazon-specific takeaways from our partner lunch: 1) AWS Fargate is an important announcement this year.** According to the partners, Amazon's managing multiple Alexa devices at work. The company is partnering with Cisco, SAP SuccessFactors, Microsoft and more to bring seamless integration between Alexa and the services provided by those companies. Using Alexa, business users can now book meetings, start a meeting, dial in a number, etc. Dr. Vogels noted that Wynn (in Las Vegas) is planning to deploy echoes in all its hotel rooms. Alexa for Business can also be integrated into third party devices such as music systems, home devices, etc.

- **AWS reInvent Twitter Discussions.** We analyzed ~22k tweets across ~11k unique accounts coming out of AWS re:Invent and note that Lambda (+Serverless), EC2, IoT, VPC, Fargate, Deep Learning, and DeepLens were some of the top mentioned topics at the event. Please see Figure 1 below.

Figure 1: Analysis of Tweets out of AWSReInvent 2017



Source: Twitter.