The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXIX, NO. 3 • JUNE 2024

Does Alternative Data Improve Financial Forecasting? The Horizon Effect

OLIVIER DESSAINT, THIERRY FOUCAULT, and LAURENT FRESARD*

ABSTRACT

Existing research suggests that alternative data are mainly informative about short-term future outcomes. We show theoretically that the availability of short-term-oriented data can induce forecasters to optimally shift their attention from the long term to the short term because it reduces the cost of obtaining short-term information. Consequently, the informativeness of their long-term forecasts decreases, even though the informativeness of their short-term forecasts increases. We test and confirm this prediction by considering how the informativeness of equity analysts' forecasts at various horizons varies over the long run and with their exposure to social media data.

THE DIGITIZATION OF INFORMATION HAS generated phenomenal growth in "alternative data" (e.g., social media, web traffic, credit card and point-of-sale, geolocation, satellite imagery, employee satisfaction ratings, etc.), transforming the way investors and information intermediaries forecast future

*Olivier Dessaint is at INSEAD. Thierry Foucault is at HEC Paris. Laurent Fresard is at the Universita della Svizzera italiana (USI) and the Swiss Finance Institute (SFI). We thank Stefan Nagel (the Editor), an anonymous Associate Editor, and two anonymous referees for helpful feedback and suggestions. We also thank Tony Cookson; Ahmed Guecioueur; Hazel Hamelin; Gerard Hoberg; Shiyang Huang; Paul Karehnke; Xinyu Liu; Adrien Matray; Randall Morck; Marina Niessner; Gordon Phillips; Thomas Philippon; Jame Russell; Eric So; Jerome Taillard; Laura Veldkamp; and participants in various conferences (AEA 2021, WFA 2021, NFA 2021, AFA 2022, Plato MI3 conference, and NBER Big Data and Securities Markets Conference) and seminars (Baruch College, CEMFI, Copenhagen Business School, CUNEFF, the Corporate Finance Webinar, ESSEC, Nova School of Business, McGill University, Neoma Business School, Norwegian School of Economics, INSEAD, Southern Methodist University, Universita della Svizzera Italiana, University of Amsterdam, University of Laval, and University of Geneva) for their comments. We thank StockTwits for providing us the data and for assistance and support. All errors are the authors' alone. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101018214). The authors have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Olivier Dessaint, INSEAD, Boulevard de Constance, 77300 Fontainebleau, France; e-mail: olivier.dessaint@insead.edu.

DOI: 10.1111/jofi.13323

outcomes.[1] Many studies find that alternative data are useful to predict firms' real outcomes or stock returns over the short term (one year or less) while there is little evidence that it is useful to predict long-term outcomes.[2] This is not surprising since many alternative data sets specifically aim to provide insights about consumers' and retail investors' current interest in firms (e.g., credit card data or social media posts about products and brands), interest that is more likely to predict firms' upcoming, rather than long-term, sales or earnings.

To the extent that alternative data are mostly "short-term-oriented" (i.e., useful to predict short-term outcomes only), we posit that the increased availability of such data has reduced the cost of predicting short-term outcomes. Intuitively, this reduction should improve the quality of forecasters' short-term forecasts. But does it also improve the quality of their long-term forecasts? We show theoretically and empirically that this is not necessarily the case.

In our theory, a forecaster must predict both the short-term and long-term earnings of a firm. She collects two types of information: (i) "short-term information" relevant to forecast short- and long-term earnings (e.g., information on the firm's assets in place), and (ii) "long-term information" relevant only to forecast long-term earnings (e.g., information about the firm's growth options). The forecaster optimally allocates effort to these two tasks to minimize the weighted sum of her average short- and long-term squared forecast errors plus the cost of effort. Importantly, she bears a multitasking cost (due, for instance, to cognitive constraints): increasing the effort allocated to one task makes effort for the other task costlier.

In this setting, we show that a drop in the cost of obtaining short-term information induces the forecaster to optimally substitute effort away from collecting long-term information. This shift in information collection makes her short-term forecasts more informative but can reduce the informativeness of her long-term forecasts, in particular when the correlation between short- and long-term earnings is low or the cost of multitasking is high. Our theory therefore highlights a downside to the availability of short-term oriented data: by altering the trade-off forecasters face in allocating their effort between the collection of short- and lon'g-term information, it can reduce the informativeness of their long-term forecasts.

We test this prediction using the forecasts of sell-side equity analysts, who routinely forecast earnings at short and long horizons considering all

[1] According to the website www.alternativedata.org, the amount invested by buy-side investors in alternative data was close to $2 billion in 2020. For implications for investors and analysts, see for example, "Demystifying Alternative Data," Greenwich Associates, 2019 or "How Investment Analysts Became Data Miners," *Financial Times*, November 28, 2019.

[2] In Table IA.I in the Internet Appendix, we summarize the results of 26 academic studies that contain results regarding the predictive power of alternative data sets for real outcomes and/or stock returns from 2000 to 2021. None reports predictability of alternative data for real outcomes or stock returns at horizons greater than one year. Data sets covered by these studies include, for instance, social media posts, product reviews, employee reviews, online customers' activity, and satellite images. The Internet Appendix may be found in the online version of this article.

relevant information, including from alternative data sources.[3] We measure the informativeness of an analyst's forecasts at a given time for the horizon h as the R^2 of a regression of realized earnings at this horizon (across the firms she covers) on her forecasts of these earnings. The higher is the R^2 for horizon h, the smaller is the residual uncertainty about firms' earnings at this horizon after observing the analyst's forecasts. A higher R^2 at horizon h thus means that the analyst's forecasts at this horizon are more informative (if $R^2 = 1$, observing the analyst's forecasts removes all uncertainty about earnings at horizon h).

Using earnings forecasts from the Institutional Brokers' Estimate System (I/B/E/S), we calculate the R^2 of each U.S. analyst every day between 1983 and 2017 for all possible horizons, ranging from one day to five years. We obtain a sample of more than 65 million analyst-day-horizon observations. As one would expect, 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 the horizon. Thus, the term structure of forecasts' informativeness (i.e., the relationship between R^2 and the forecasting horizon) is downward-sloping. Our theory predicts that more short-term oriented data should increase the R^2 of analysts' short-term forecasts, but can decrease the R^2 of their long-term forecasts, thereby leading to a steepening of the term structure.

We first test whether the long-run evolution of forecasts' informativeness is consistent with this prediction. Figure 1 shows that the number of alternative data sets has been growing since the 1990s. If this growth represents mainly an expansion of short-term-oriented data, then our theory implies that, over time, the informativeness of analysts' short-term forecasts should have improved while the informativeness of their long-term forecasts should have decreased. We find that these two opposite trends 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 has decreased at the five-year horizon from about 40% to 30%. We also show that the "slope" of the term structure has become steeper over time, a trend that has accelerated since 2005. This pattern applies to most industries. Our theory implies that it should be more pronounced in those industries covered by analysts using more alternative data. This is what we find using references to the use of alternative data in analysts' reports.

These findings are consistent with our theory but do not constitute causal evidence, factors other than the growth of short-term-oriented data could explain the long-run evolution of the term structure of forecast informativeness and its variation across industries. To provide a tighter test, we exploit the introduction of the social media platform StockTwits, which expands the amount of information available to analysts. On StockTwits, investors can share

[3] See Chi, Hwang, and Zheng (2021) for evidence that analysts use alternative data and Section IA.XVII in the Internet Appendix for an example.

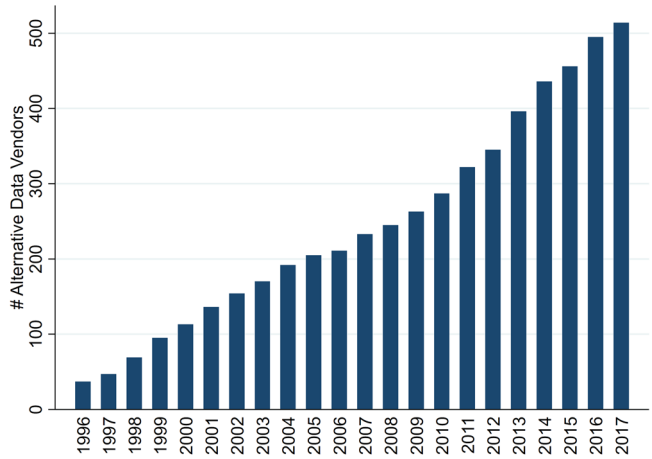

Figure 1. Estimated number of alternative data vendors. This figure shows the estimated number of alternative data vendors by year based on the J.P. Morgan 2019 *Alternative Data Handbook*. The J.P. Morgan handbook (Kolanovic and Smith (2019)) comes with a directory of alternative data vendors as of 2019 with a hyperlink to their URL domain (see pp. 204–255). We use this hyperlink, and Internet Archive Wayback Machine (https://web.archive.org/), to retrieve the entire web page history of each vendor saved by the Wayback Machine since 1996. We fill gaps in the time series of records when gaps are smaller than five years, and drop records before the gap otherwise. We estimate the number of vendors every year by counting the vendors with one or more pages saved on the Internet Archive in that year. (Color figure can be viewed at wileyonlinelibrary.com)

information and opinions about individual firms.[4] This platform is well suited for our analysis because the data it generates (i) cover almost every firm (in contrast to many other alternative data sets, whose coverage is more specialized), (ii) are short-term-oriented (we show that it predicts firms' outcomes up to one year ahead but not beyond), and (iii) are used by analysts (we provide evidence thereof). Moreover, StockTwits was introduced in 2009 and expanded progressively with different levels of intensity across firms. This feature enables us to estimate the effect of greater exposure to alternative (social media) data.

We measure analysts' exposure to data specifically generated on StockTwits (i.e., not available to analysts from traditional sources) using two complementary approaches. First, we use the daily average number of users who have the firms covered by a given analyst on their "watchlist" (i.e., the list of firms they follow). Since users rarely modify their watchlist after registering on StockTwits, a firm's watchlist changes because new users register and enter the platform. Therefore, variation in the number of users on the watchlists of

[4] Several recent academic papers use data from StockTwits to capture, for instance, divergence of investors' opinions (Cookson and Niessner (2020), or Giannini, Irvine, and Shu (2019)), the political orientation of investors' beliefs (Cookson, Engelberg, and Mullins (2020)), or selective exposure to confirmatory information (Cookson, Engelberg, and Mullins (2022)). In contrast, we use StockTwits to capture variation in the availability of short-term-oriented data.

covered firms mostly reflects the overall expansion of StockTwits and not the arrival of information from other sources. Second, we use the average number of "hypothetical" messages about the covered firms over the last 30 days. Hypothetical messages correspond to the total number of messages on StockTwits multiplied by a firm's average share of total messages. Because this share is constant, the number of hypothetical messages about a firm does not change with the arrival of firm-specific information from other sources. 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 subsample over the 2005 to 2017 period.

We find that greater exposure to data generated on StockTwits 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. This steepening is more pronounced for analysts following more firms (i.e., those for whom the cost of multitasking is plausibly higher), and for analysts following firms whose earnings are less autocorrelated, as our theory predicts.

In sum, our empirical findings based on StockTwits data are consistent with our prediction that the increased availability of short-term-oriented data can reduce the informativeness of long-term forecasts. This conclusion does not necessarily apply to any type of alternative data set. According to our theory, it should only apply to short-term-oriented data sets. For data sets containing information about long-term outcomes, our predictions are reversed: the introduction of such data should make long-term forecasts more informative, possibly at the expense of the informativeness of short-term forecasts. Identifying variation in the availability of long-term-oriented data would offer another way to test the economic forces at play in our model.

The rest of the paper is organized as follows. Section I positions our contribution in the related literature. In Section II, we present our model and derive our main prediction. Section III presents the data used in our tests and our measure of analysts' forecast informativeness. Sections IV and V report the findings of our main tests. Section VI concludes. All derivations and definitions of the variables used in our tests are reported in the Appendices.

I. Contribution to the Literature

Our results add to the growing literature on 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, and study the implications for the informativeness of asset prices (Dugast and Foucault (2018), Farboodi and Veldkamp (2020)), market efficiency (Martin and Nagel (2022)), firms' growth rates (Begeneau, Farboodi, and Veldkamp (2018)), information acquisition choices of asset managers (Abis (2018), Dugast and Foucault (2022)), the pricing of

information by data vendors (Huang, Xiong, and Yang (2022)), or financial inclusion (Mihet (2020)).

Following this literature, we assume that the growth of short-term-oriented data reduces the cost of information acquisition. We consider the possibility however that this reduction is *heterogeneous* across forecasting horizons. 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- and long-term information. This trade-off is relevant because most financial decisions require forecasting outcomes that occur at different dates in the future.[5]

A growing body of research explores the potential of various alternative data sources in predicting returns or other outcomes, such as earnings. For instance, Green et al. (2019) demonstrate that employee reviews sourced from Glassdoor can be used to predict stock returns. The literature suggests that alternative data contain valuable information, although the utility of such data lies primarily in forecasting short-term outcomes (see Table IA.I for an overview of existing evidence). In this paper, we show that this can induce forecasters to reallocate effort from forecasting long-term outcomes to forecasting short-term outcomes.

Our focus on the informativeness of financial forecasts distinguishes our study from existing papers that analyze the effects of digitization on the informativeness of order flows and asset prices. Using the digitization of firms' regulatory filings (Gao and Huang (2020)), the availability of satellite images of retailers' parking lots (Zhu (2019)), or variation in the volume of data generated by financial blog posts (Grennan and Michaely (2020)), these papers show that increased digitization strengthens the informativeness of order flow and stock prices at short horizons (up to one year). Relatedly, Bai, Philippon, and Savov (2016) and Farboodi et al. (2021) examine whether long-run reductions in information processing costs have changed the informativeness of stock prices. They report mixed effects, with improvements for some firms but deteriorations for others.

These findings are not directly comparable to ours because the informativeness of stock prices is distinct from the informativeness of analyst forecasts. Nevertheless, our findings are not inconsistent with those on the evolution of stock price informativeness. Intuitively, as stock prices are the sum of discounted forecasted cash flows at all horizons, their informativeness about firms' cash flows at specific future dates is a weighted average of the informativeness of both short- and long-term forecasts. Thus, the net effect of an increase in the informativeness of short-term forecasts and a decrease in the informativeness of long-term forecasts on stock price informativeness can be

[5] Dugast and Foucault (2018) show 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 notion of short term and long term used in our paper, namely, the time elapsed *until* the realization of a payoff.

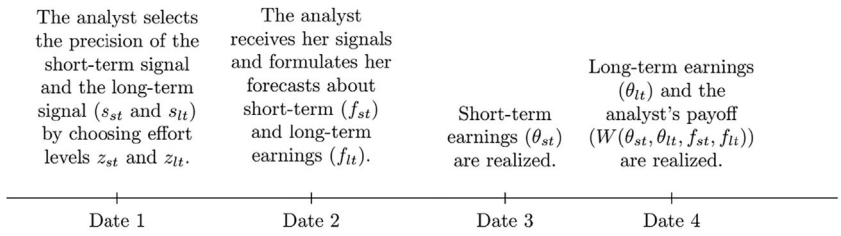

Figure 2. Timeline of the model.

positive or negative depending on the environment (e.g., the level of discount rates or the maturity of cash flows).

Finally, our paper also contributes to the literature that studies how analysts form their forecasts. As explained in Section IV.B, our measure of analyst forecast informativeness is similar to that of Hilary and Hsu (2013). To our knowledge however, our finding of a downward-sloping term structure of forecasts' informativeness and its steepening over time is novel. More importantly, we relate this term structure to the type of data available to analysts and their allocation of efforts between different tasks. Our findings therefore 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- and long-term forecasts (Bandyopadhyay, Brown, and Richardson (1995) or Mest and Plummer (1999)), and how progress in information technologies affects the organization and output of the financial analysis industry (Chi, Hwang, and Zheng (2021), Gerken and Painter (2022), van Binsbergen, Han, and Lopez-Lira (2022), or Grennan and Michaely (2021)).

II. Model

A. *The Forecasting Problem*

The model features one forecaster (the "analyst") and one firm. Figure 2 presents the timeline. The firm generates earnings, θ_{st} at date 2 (the short term) and θ_{lt} at date 3 (the long term) according to

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

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. As can be seen, long-term earnings are the sum of two components: (i) the *common component*, $\beta\theta_{st}$, which is generated, for instance, by assets in place, and (ii) *the unique component*, e_{lt}, which is generated, for instance, by growth opportunities.

At date 1, the analyst formulates forecasts about θ_{st} and θ_{lt}, denoted by f_{st} (the short-term forecast) and f_{lt} (the long-term forecast), respectively. Her payoff, $W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st})$, 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 II.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 long-term earnings,

$$s_{st} = \theta_{st} + \tau_{st}(z_{st})^{-\frac{1}{2}}\epsilon_{st}, \quad s_{lt} = e_{lt} + \tau_{lt}(z_{lt})^{-\frac{1}{2}}\epsilon_{lt}, \tag{3}$$

where $\epsilon_{st} \sim \mathcal{N}(0, \sigma_{st}^2)$, $\epsilon_{lt} \sim \mathcal{N}(0, \sigma_e^2)$, and $\tau_h(z_h) = \frac{\psi_h z_h}{1-\psi_h z_h}$, and where $z_h < \psi_h^{-1}$ (chosen at date 0; see below) is the level of effort that the analyst exerts to collect and process information at horizon $h \in \{st, lt\}$. The errors (ϵ_h s) in the analyst's signals are independent of each other and of all other variables in the model. Given this specification, we have

$$\mathsf{var}(\theta_{st}|s_{st}) = \sigma_{st}^2(1 - \psi_{st}z_{st}), \quad \text{and} \quad \mathsf{var}(e_{lt}|s_{lt}) = \sigma_e^2(1 - \psi_{lt}z_{lt}). \tag{4}$$

By exerting effort (z_{st}) to collect short-term information, the analyst improves the precision of her signals ($\tau_h(z_h)$ increases with z_h), which reduces uncertainty about the common component of future earnings ($\mathsf{var}(\theta_{st}|s_{st})$ decreases with z_{st}). In contrast, by exerting effort to collect long-term information, the analyst reduces uncertainty about the unique component of long-term earnings only. Effort to reduce uncertainty about one component does not affect the uncertainty about the other component because these efforts require distinct tasks (the unique component is not correlated with the common component). Each unit of effort to obtain information about one component reduces prior uncertainty about this component by a constant fraction ψ_h. Thus, ψ_h measures the "informational return" of effort at horizon h.

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

$$\begin{aligned}\overline{W}(f_{st}, f_{lt}; s_{st}, s_{lt}) &\equiv \mathsf{E}(W(\theta_{st}, \theta_{lt}, f_{lt}, f_{st})|s_{st}, s_{lt}) \\ &= \omega - \gamma\mathsf{E}((f_{st} - \theta_{st})^2|s_{st}, s_{lt}) - (1-\gamma)\mathsf{E}((f_{lt} - \theta_{lt})^2|s_{st}, s_{lt}).\end{aligned} \tag{5}$$

The analyst chooses her forecasts $\{f_{st}^*, f_{lt}^*\}$ to maximize $\overline{W}(f_{st}, f_{lt}; s_{st}, s_{lt})$. Thus,

$$f_{st}^* = \mathsf{E}(\theta_{st}|s_{st}), \quad f_{lt}^* = \mathsf{E}(\theta_{lt}|s_{st}, s_{lt}) \tag{6}$$

(see Appendix C for derivations).

Substituting equation (6) into (5), we obtain that the analyst's unconditional (date 0) expected payoff is

$$\mathsf{E}(\overline{W}(f_{st}^*, f_{lt}^*; s_{st}, s_{lt})) = \omega - q(\beta, \gamma)\mathsf{var}(\theta_{st}|s_{st}) - (1-\gamma)\mathsf{var}(e_{lt}|s_{lt}), \tag{7}$$

where $q(\beta, \gamma) \equiv \gamma + (1-\gamma)\beta^2$. The analyst's *expected* payoff increases when her signals are more precise (i.e., $\mathrm{var}(\theta_{st}|s_{st})$ and $\mathrm{var}(e_{lt}|s_{lt})$ are smaller) because more precise signals reduce her average squared forecast errors.

Exerting effort is costly. The analyst's 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. Following 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 multitasking is costly, $c > 0$. For instance, if the analyst has exerted a high level of effort to collect, say, short-term information, then it is more costly for the analyst to exert even more effort, be it to reduce the uncertainty about the common component of long- and short-term earnings ($a > 0$) or uncertainty about the unique component of long-term earnings ($c > 0$).[6]

The analyst chooses her effort levels 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

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

In choosing effort levels, the analyst trades off the precision of her signals and the cost of effort. We obtain the following result.[7]

PROPOSITION 1: *Suppose $c < \bar{c}(\beta, \gamma, a, b, \psi_{st}, \psi_{lt})$ (where $\bar{c}$ is defined in the proof of the proposition), $\psi_{st} < (2a/\sigma_{st}^2 q(\beta, \gamma))^{\frac{1}{2}}$, and $\psi_{lt} < (2b/\sigma_e^2(1-\gamma))^{\frac{1}{2}}$. Under these conditions, the analyst's optimal levels of effort in producing information at date 0, z_{st}^* and z_{lt}^*, are interior (i.e., $0 < z_h^* < (\psi_h)^{-1}$) and given by*

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

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 the short-term signal.

[6] In Goldstein and Yang (2015), the payoff of an asset is the sum of three components and investors can acquire information about the first component, the second component, 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 $c > 0$.

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

The analyst responds by exerting more effort to obtain short-term information. This response is optimal but it raises the marginal cost of exerting effort to improve the precision of the long-term signal because multitasking is costly ($c > 0$). Consequently, the analyst also optimally reduces the effort she allocates to this task. This mechanism can work via either a decrease 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 effort allocated to each task (see Section II.C).

B. Short-Term-Oriented Data and Forecasts' Informativeness

As explained in the introduction, existing studies suggest that alternative data are mainly short-term-oriented. Thus, the availability of such data reduces the cost of obtaining short-term information, a in the model. According to Proposition 1, such a decrease should lead the analyst to exert more effort to improve the precision of her short-term signals at the expense of the precision of her long-term signals. Testing this prediction directly is difficult because analysts' effort choices are not directly observable. However, as shown in Corollary 1, one can use the informativeness of analysts' forecasts to test this implication of Proposition 1. Intuitively, the analyst's forecast at horizon h is more informative if, after observing it, residual uncertainty about future earnings ($\mathsf{var}(\theta_h \mid f_h^*)$) relative to prior uncertainty ($\mathsf{var}(\theta_h)$) is smaller. We therefore define the informativeness of the analyst forecast at horizon h, denoted by R_h^2, as

$$R_h^2 \equiv \frac{\mathsf{var}(\theta_h) - \mathsf{var}(\theta_h \mid f_h^*)}{Var(\theta_h)} \quad \text{for} \quad h \in \{st, lt\}. \tag{11}$$

The higher is R_h^2, the greater is the informativeness of the analyst's forecast at horizon h. If the analyst's forecast is fully informative, then $R_h^2 = 1$ since in this case $\mathsf{var}(\theta_h \mid f_h^*) = 0$. In contrast, if the analyst's forecast is fully uninformative, then $R_h^2 = 0$ since in this case $\mathsf{var}(\theta_h \mid f_h^*) = \mathsf{var}(\theta_h)$.[8]

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

$$R_{st}^2 = \psi_{st} z_{st}^*, \quad \text{and} \tag{12}$$

$$R_{lt}^2 = (1 - \rho^2)\psi_{lt} z_{lt}^* + \rho^2 \psi_{st} z_{st}^* \tag{13}$$

(see Appendix C for the algebra), where $\rho = (\frac{\beta^2 \sigma_{st}^2}{\beta^2 \sigma_{st}^2 + \sigma_e^2})^{\frac{1}{2}}$ is the correlation between the firm's short- and long-term earnings. The informativeness of the

[8] Campbell and Thompson (2008) show that the expected utility gain of observing a predictor of asset returns for a mean-variance investor increases with the R^2 of a regression of returns on the predictor. Thus, R_h^2 is not only an intuitive way to measure the informativeness of a forecast but also a measure of the value of this forecast for a mean-variance investor (see Thesmar and de Silva (2021, Section 2) for a similar point in the context of analyst forecasting).

analyst's short-term forecast depends only on her optimal level of effort, z^*_{st}, to collect information about the common component of the firm's future earnings and is increasing 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 long-term earnings (if $\rho > 0$).

COROLLARY 1: *Let* $\bar{\rho} = (\frac{c\psi_{lt}}{2b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$. *A decrease in the marginal cost of producing the short-term signal, a, has:*

1. *A negative effect on the informativeness of the long-term forecast if* $\rho < \bar{\rho}$ *and a positive effect otherwise (i.e., when* $\rho \geq \bar{\rho}$*).*
2. *A positive effect on the informativeness of the short-term forecast.*
3. *A negative effect on* $\Delta = R^2_{lt} - R^2_{st}$*, the difference between the informativeness of the long- and short-term forecasts.*

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- and short-term earnings, ρ, is smaller than a threshold $\bar{\rho}$. This condition is satisfied for firms with a low ρ and/or for analysts with a high cost of multitasking, c (as $\bar{\rho}$ increases with c).[9] In this case, the informativeness of the long-term forecast declines with the cost of producing short-term information (part 1 of Corollary 1). In contrast, the informativeness of the short-term forecast always improves (part 2).

These differential effects constitute our main testable implication. Insofar as short-term-oriented data increase 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 expanded availability should coincide with an increase in the informativeness of short-term forecasts and a *decrease* in the informativeness of long-term forecasts for analysts following firms with a relatively low autocorrelation of earnings (i.e., low ρ) or facing a high cost of multitasking (c). Moreover, the last part of Corollary 1 shows that, in all cases, the availability of short-term-oriented data should increase the difference between the informativeness of the long- and short-term forecasts (Δ). Empirically, this difference (the "slope" of the term structure of forecast informativeness) is negative (see Section III.C). Thus, the model implies that a decrease in the cost of producing short-term information should make it more negative, that is, should steepen the slope.

[9] 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 Appendix C, it can be checked that the set of parameter values (e.g., for c and β) such that $c < \bar{c}$ and $\rho < \bar{\rho}$ is nonempty.

Remark: Equations (12) and (13) imply that $R^2_{st} > R^2_{lt}$ if and only if $\psi_{st} z^*_{st} > \psi_{lt} z^*_{lt}$. That is, the informativeness of the short-term forecast is higher only if the analyst exerts sufficiently high effort (z^*_{st}) in producing short-term information relative to the effort she exerts for producing long-term information (z^*_{lt}). This is the case (see equation (10)) if (i) a is low enough relative to b, or (ii) the analyst's payoff depends more on her short-term forecasting error (i.e., γ is large). Intuitively, both conditions are likely to hold in reality. We find that the informativeness of short-term forecasts is indeed larger than the informativeness of long-term forecasts (see Figure 3 and Table I). However, none of our predictions depends on this feature of the data.

C. Discussion and Interpretation

Analysts' Objective Function. As assumed in our model (see equation (2)), analysts care about their forecast errors because such errors affect their career outcomes. For instance, Hong and Kacperczyk (2010) and Harford et al. (2019) show that analysts with smaller forecast errors are more likely to be ranked "All Star" analysts or to be promoted. This relationship might be direct, for instance, when 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 firm based on her forecasts.

In reality, analysts issue long-term forecasts less frequently than short-term forecasts (see Section III.A). However, this fact does not imply that they care only about 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 therefore depends on both the quality 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, Shane, and Yang (2012)). Moreover, revisions of long-term forecasts lead to strong market reactions (Chen, Da, and Zhao (2013), Da and Warachka (2011), or Copeland, Dogloff, and Moel (2004)), suggesting that long-term forecasts matter for investors, which supports our assumption that long-term forecasting errors matter for analysts' careers. (i.e., $\gamma < 1$ in (equation (2)).

Splitting Tasks. When the cost of obtaining short-term information drops, the analyst reallocates effort to collecting short-term information. This behavior is optimal because it saves on the cost of multitasking. However, it can increase the analyst's long-term forecasting error. One may then wonder whether the analyst (or her employer) would be better off by dividing the tasks of forecasting short- and long-term earnings between two agents. In Section IA.I of the Internet Appendix, we show that this is not the case when $c \leq \frac{4C_0}{q(\beta,\gamma)(1-\gamma)\psi_{st}\psi_{lt}\sigma_e^2\sigma_{st}^2}$. Indeed, under this condition, the increase in fixed costs of information production (each agent bears the fixed cost C_0 of collecting

information to understand the firm's business) cancels out savings on the cost of multitasking. As we also discuss in Section IA.I of the Internet Appendix, agency frictions can also explain why splitting the tasks of forecasting short- and long-term earnings between two agents is not optimal.

Alternative Interpretation. Instead of reducing the cost of obtaining short-term information, the availability of short-term-oriented data can increase the informational return on effort for obtaining short-term information, ψ_{st} (e.g., because short-term-oriented data lend themselves more easily to quantitative analysis). Our main prediction in this case is unchanged (see Section IA.II in the Internet Appendix). In particular, if $\rho < (\frac{c\psi_{lt}}{4b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$ (a condition qualitatively similar to that in Corollary 1), then an increase in ψ_{st} improves the informativeness of the analyst's short-term forecast and reduces that of her long-term forecast. What matters for our prediction therefore is that the increased availability of short-term-oriented data raises the marginal net benefit of effort for obtaining short-term information.

III. Forecasts' Informativeness: Data and Measurement

In this section, we first explain how we measure the informativeness of earnings forecasts issued by analysts. We then present summary statistics.

A. Analysts' Earnings Forecasts and Realizations

We retrieve analyst forecasts of earnings per share (EPS) and net income (in U.S. dollars) from the I/B/E/S Detail History File (Adjusted and Unadjusted) at different horizons (up to five years). We exclude quarterly and semiannual earnings forecasts, and we retain annual earnings forecasts associated with a well-defined fiscal period.[10] We eliminate forecasts with missing announcement dates, analyst codes, or broker codes. 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 the Center of Research in Security Prices (CRSP), and forecasts for firms with missing information on stock price, or number of shares, and forecasts with share code other than 10, 11, or 12.

We use net income forecasts as our main measure of "earnings" forecasts.[11] We match earnings forecasts to realized earnings reported in the I/B/E/S

[10] We identify forecasts for different fiscal years using I/B/E/S item "fpi" and retain forecasts with $fpi = 1, 2, 3, 4, 5$. Section IA.V of the Internet Appendix explains why we do not consider long-term growth forecasts.

[11] If an analyst simultaneously issues a net income forecast 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 in which we observe both an EPS forecast 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 to the I/B/E/S-adjusted actual EPS (see Section IA.IV in the Internet Appendix).

Actual File. By default, we use 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 and the number of shares from the CRSP otherwise. Finally, 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 correspond to a fiscal year ending between 1983 to 2017, (iii) forecasts are 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 a sample of 9,129,282 unique forecasts and realizations by analyst-firm-date-horizon.

The sample contains 4,259,465 million forecasts with horizon less than one year and 1,260,796 million with a horizon greater than two years (including 102,431 beyond four years), where horizon is the number of days between the forecast date and the earnings report date, divided by 365. Two factors explain why there are more short-term than long-term unique forecasts. First, for inclusion in our sample, the earnings realization must be nonmissing. As the horizon increases, some firms become inactive before we observe the corresponding earnings. Second, 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 tend to occur after annual earnings announcements.[12] Arguably, analysts and firms for which we observe long-term forecasts might be different. To mitigate concerns about selection, below we verify that our results hold for analysts who release both short- and long-term forecasts and for firms with both types of forecasts.

B. Measuring Forecast Informativeness

In the model, we define the analyst's forecast informativeness at horizon h as R_h^2, the fraction of the variance of the firm's earnings at horizon h that is explained by the forecast of these earnings (see equation (11)). Empirically, we estimate $R_{i,t,h}^2$ for a given analyst i on day t for horizon h as the R^2 of the regression

$$e_j = k_0 + k_1 f_j + \nu_j, \tag{14}$$

where j indexes all firms covered by analyst i at time t with an available forecast at horizon h, and where f_j and e_j are, respectively, the forecasted and realized earnings for firm j normalized by total assets. As we explain in Section II.B, a higher $R_{i,t,h}^2$ indicates that analyst i's forecast at horizon h on day

[12] As we discuss in Section II.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.

t is more informative in the sense that observing her forecasts reduces the residual uncertainty ($\text{var}(\nu_j)$) about earnings at this horizon by a larger amount relative to prior uncertainty ($\text{var}(e_j)$).

This measure has several advantages. First, according to the model, it is a direct measure of analysts' effort to produce information at a given horizon (see equations (12) and (13)). Second, R^2 is a normalized measure scaled between zero (the analyst's forecast is only noise) and one (the analyst has perfect foresight). This normalization allows for meaningful comparisons of forecast informativeness across analysts at a given time or over time for a given analyst. A closely related measure is the variance of earnings at horizon h explained by the analyst's forecast at this horizon ($\text{var}(k_0 + k_1 f_j) = \text{var}(e_j)R^2$ by definition of R^2). However, variation in this measure over time or across analysts can be due to either variation in prior uncertainty ($\text{var}(e_j)$) or variation in analysts' effort to reduce uncertainty. Our theory corresponds to the second source of variation, which is captured by R^2 (see equations (12) and (13)).

The literature considers other measures of analyst forecasts quality, namely, (i) the impact of their forecasts on stock prices (Merkley et al. (2017)) or (ii) their absolute (or squared) forecast error (often called "accuracy"). For our purposes, these measures are problematic. First, 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 since one cannot disentangle the contribution of each forecast to the price reaction. Second, the absolute (or squared) error of an analyst can be large and yet her forecast be informative if the error is always the same, that is, if the analyst is systematically biased (for example, due to conflicts of interest (Hong and Kacperczyk (2010))). Yet, as Hilary and Hsu (2013) stresses, accuracy and informativeness are not the same.[13] To address this issue, Hilary and Hsu (2013) propose that the inverse of the variance of analysts' unexpected forecast errors be used to measure forecasts informativeness. This measure is very similar to our measure.[14] Like their measure, R^2 is not affected by the average level of the analyst's bias (see Section IA.VI in the Internet Appendix).

Appendix A describes the detailed procedure to compute R^2 for an analyst on a given day for a given horizon. We apply this procedure for all analysts at all dates between January 1, 1983 and December 31, 2017 and for all possible horizons between one day and five years. Our final sample contains 65,889,122

[13] Hilary and Hsu (2013) provide the following example. Consider two analysts: A and B. Analyst A's forecasts are consistently three cents below realized earnings while Analyst B's forecasts are two cents above (below) realized earnings half of the time. Analyst A's squared forecast error is larger (9 cents vs. 4 cents) but it is more informative because after adding three cents to her forecast, one has a perfect estimate of future earnings. They conclude that: *"[...] the usefulness of analysts' forecasts should not be based on forecasts' stated accuracy (their absolute distance from realized earnings), but rather on forecasts' informativeness"* (p. 272).

[14] In our setting, an analyst's expected forecast error at horizon h is $\text{E}(e_j - f_j) = k_0 + (k_1 - 1)f_j$ and therefore the unexpected forecast error is $e_j - f_j - (k_0 + (k_1 - 1)f_j) = \nu_j$. The inverse of the variance of unexpected forecast errors is therefore $\text{var}(\nu_j)^{-1} = (\text{var}(e_j)(1 - R^2))^{-1}$. Thus, when an analyst's forecast informativeness (R^2) is high according to our measure, it is also high in the sense of Hilary and Hsu (2013).

Table I
R^2 Summary Statistics

This table presents descriptive statistics for the main analyst-day-horizon variables over the 1983 to 2017 period. 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 firm observations used to estimate equation (14). We present statistics for the whole sample, as well as subsamples including observations in different annual forecasting horizon ranges. Variable definitions are in Appendix B.

	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

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. Summary Statistics and Stylized Facts

Table I presents summary statistics for R^2.[15] Over the 1983 to 2017 period, analysts' earnings forecasts explain on average 68.01% of the variation in realized earnings across the firms they cover. The average horizon of their forecasts is 1.11 years, and they cover 8.12 firms. The average R^2 decreases with horizon, from 79.60% for horizons shorter than one year to 59.21% for horizons between one and two years, 49.37% for horizons between two and three,

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

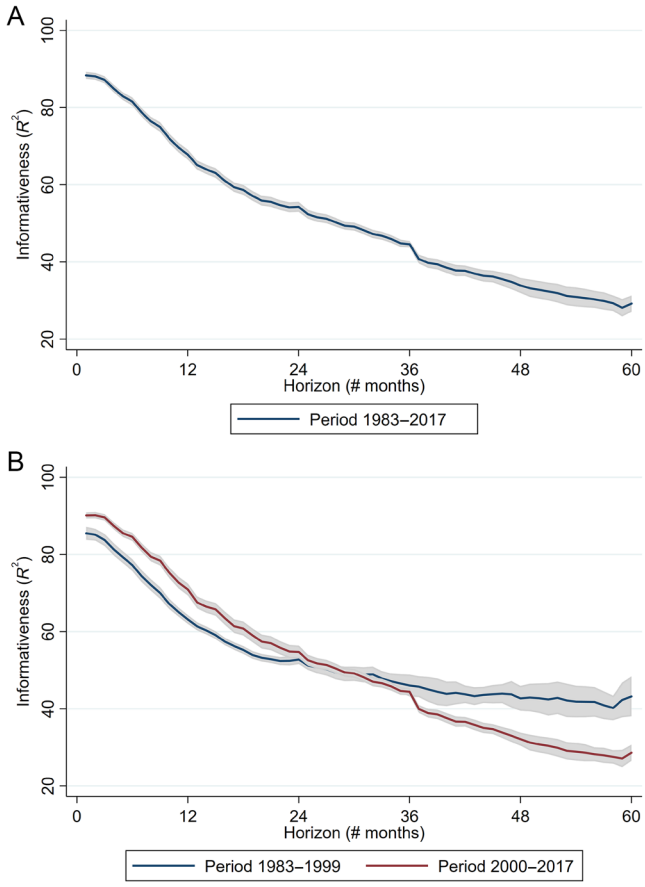

Figure 3. The term structure of analysts' forecast informativeness. This figure displays the term structure of analysts' forecast informativeness. Each graph shows the means of analyst-level $R^2_{i,t,h}$ over all analysts and dates, for fixed h values expressed in number of months (displayed on the 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 to 2017 (Panel A), split into two subperiods (Panel B). The shaded gray area corresponds to a 90% confidence interval, obtained with standard errors clustered by forecasting fiscal period. (Color figure can be viewed at wileyonlinelibrary.com)

37.62% between three and four years, and 31.18% for horizons beyond four years. We refer to the relationship between R^2 and forecasting horizon h as the term structure of forecasts' informativeness, or simply the "term structure."

To characterize the shape of this term structure, we plot the averages of analyst-level $R^2_{i,t,h}$ across all analysts and days by horizon h (expressed in months). Figure 3, Panel A, confirms that the term structure is downward-sloping. The slope of this term structure (Δ in our theory), estimated by regressing the averages of R^2 by h on h, is negative and equal to -1 (t-statistic $= -24$). Its intercept is 81 (t-statistic $= 54$). This linear approximation implies

that informativeness deteriorates by 1 percentage point for every one-month increase in the horizon, that is 12 percentage points per year.

Corollary 1 implies that greater exposure to short-term-oriented data increases R^2 for low values of h, but possibly decreases R^2 for high values of h, thereby steepening the slope of the term structure. We therefore use two approaches to test Corollary 1, one that focuses on the level of R^2 by horizon and studies variation in R^2 separately for fixed values of h, and one that focuses on the slope of the term structure, which we estimated with a linear approximation.

IV. Long-Run Evolution

The number of alternative data sets has increased steadily over time, with an acceleration in recent years (see Figure 1). If these data sets are mainly short-term-oriented, Corollary 1 implies that (i) the informativeness of analysts' short-term forecasts should have improved over time while that of their long-term forecasts should have deteriorated, and (ii) these opposite trends being more pronounced among analysts covering firms for which more alternative data sets have become available. We test these implications in this section.

A. Forecast Informativeness by Horizon

We begin by illustrating the evolution of the term structure. To do so, we split the sample into two periods of equal length (1983 to 1999 and 2000 to 2017) and compare the average term structure over each period. Panel B of Figure 3 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 2000.[16] To verify that this steepening corresponds to a general trend, we compute and plot the averages of $R^2_{i,t,h}$ by year, separately for short-term ($h < 1$) and long-term ($h \geq 2$) forecasts. Figure 4 confirms the presence of two opposing trends: an improvement in R^2 for short-term forecasts, and a deterioration in R^2 for long-term forecasts.

To formally test whether these opposite trends are statistically significant, we regress $R^2_{i,t,h}$ on a year counter variable by horizon subsample. This counter is set to zero before 1989 and increases by one every year after. We divide this variable by the number of years between 1990 and 2017 so that the estimated coefficient corresponds to the cumulative change in R^2 over the 1990 to 2017 period.[17] Results are reported in Table II and confirm the patterns in Figure 4. For horizons shorter than one and two years, the average R^2 has increased by

[16] For example, Srinidhi, Leung, and Jaggi (2009) document an improvement in the precision of the idiosyncratic information component in short-term forecasts in the two years following regulation Fair Disclosure (RegFD) in 2000 (compared to two years prior), whereas that of long-term forecasts declined.

[17] In this test and in the rest of the paper, we cluster standard errors by forecasted fiscal period, except in Tables III and IV and Figure 7, because observations are not available by forecasted fiscal period.

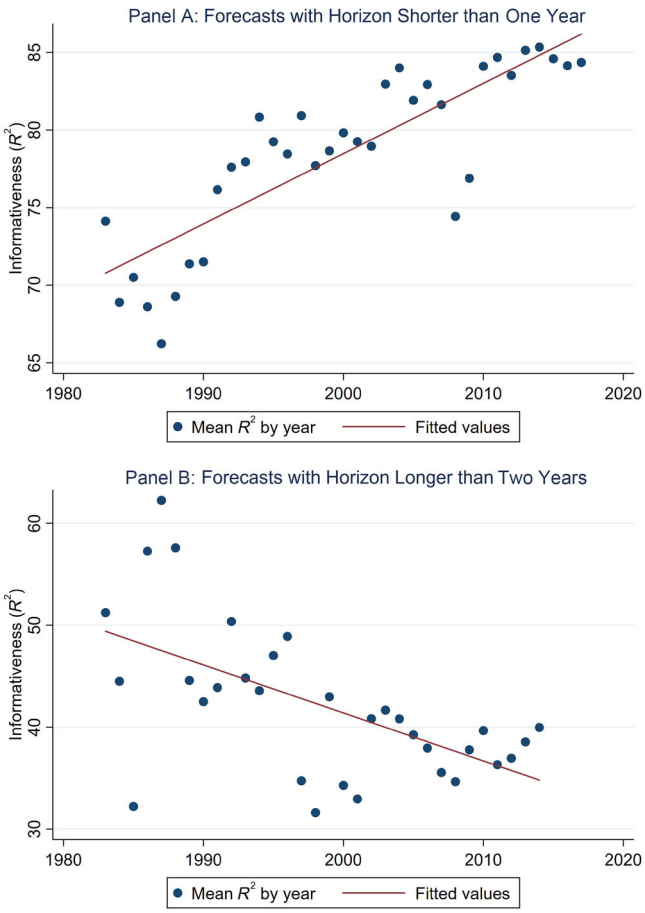

Figure 4. Short-term versus long-term forecast informativeness by year. This figure shows the means of analyst-level $R^2_{i,t,h}$ over all analysts by year, separately for short-term ($h < 1$) and long-term ($h \geq 2$) forecasts. The sample period is 1983 to 2017 for short-term forecasts (Panel A), and 1983 to 2015 for long-term forecasts (Panel B). (Color figure can be viewed at wileyonlinelibrary.com)

12.7 (column (1)) and 9.9 (column (2)) percentage points, respectively. Beyond three and four years, the average R^2 has deteriorated by 12.4 (column (4)) and 21.8 (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 the ordinary least squares (OLS). Figure 5 shows the evolution of the slope estimates. The slope

Table II

Forecast Informativeness by Horizon

This table presents OLS estimates of the time trend in analysts' forecast informativeness by subsamples 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* takes the value of zero for the period 1983 to 1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative increment in R^2 over the 1990 to 2017 period. Variable definitions are in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Forecasts 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	12.7***	9.9***	2.4	−12.4***	−21.8***
	(8.82)	(7.18)	(1.38)	(−5.11)	(−5.52)
Constant (83–89)	73.6***	54.3***	47.8***	45.3***	44.5***
	(93.57)	(80.88)	(36.20)	(27.17)	(19.38)
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712

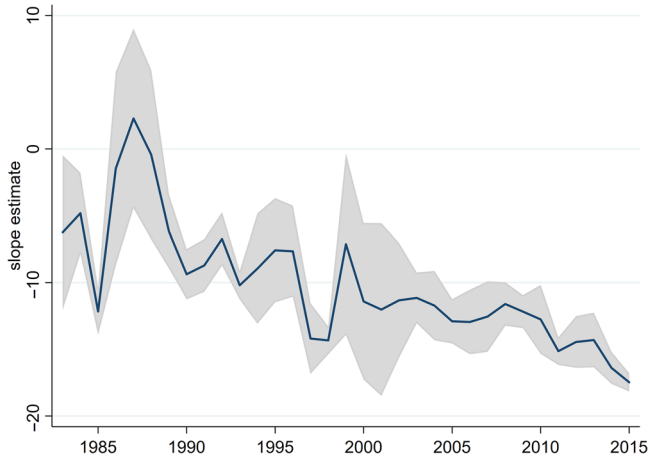

Figure 5. The slope of term structure by year. This figure shows the evolution over time of the slope of the term structure of analysts' forecast informativeness. The slope is estimated every year by linear approximation, regressing the average of R^2 by horizon h on h, separately for every calendar year. The figure plots the regression coefficient. Each slope estimate indicates 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 one year. The shaded gray area corresponds to a 90% confidence interval, obtained with standard errors clustered by forecasting fiscal period. (Color figure can be viewed at wileyonlinelibrary.com)

Table III
The Slope of the Term Structure

This table presents OLS estimates of the 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 the horizon increases by one year. 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 Fama-French 17 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. The regression coefficients on h used as estimates for the slope are winsorized at the 1% level in each tail. The slope is estimated only if there is at least one forecast with $h \geq 3$. *Year Trend* takes the value of zero for the period 1983 to 1989 and increments by one every subsequent year divided by 28 so that the regression coefficient can directly be interpreted as the cumulative change in slope over the 1990 to 2017 period. t-Statistics in parentheses are based on standard errors clustered by year. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Slope by Year	Slope by FF17-Year		Slope by Analyst-Year	
OLS:	(1)	(2)	(3)	(4)	(5)
Year Trend	−11.3***	−5.8***	−4.5***	−5.4***	−3.0**
	(−6.67)	(−5.44)	(−4.36)	(−8.11)	(−2.12)
Constant (83−89)	−5.8***	−10.1***		−11.6***	
	(−5.39)	(−18.05)		(−23.71)	
FF17 FE	–	No	Yes	–	–
Analyst FE	–	–	–	No	Yes
N	33	483	456	7,657	7,290

was around −10 until the mid-1990s, but then became steeper, especially after 2005. After this date, the slope is consistently smaller than −10. Table III confirms this pattern. Regressing the slope estimates on a year counter and a constant (as we do above), column (1) shows that the term structure steepens over time, with an average slope that shifts from −5.8 during the baseline period 1983 to 1989 to (−5.8−11.3 =) −17.1 in recent years.

The rest of Table III shows that this pattern holds when we first estimate the slope by Fama-French 17 (FF17) industry and year, and when we estimate it by analyst and year (for industry-year and 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 both within industry and within analyst, and thus that our findings are not likely driven by changes in sample composition. Results in column (5) also demonstrate that the selection of different analysts in our sample (e.g., some with and others without missing long-term forecasts) cannot drive our finding, since we observe the same change in term structure for the same analyst over time.

To further alleviate concerns regarding changes in sample composition, in Sections IA.X to IA.XIII of the Internet Appendix we show that the trend documented in Tables II and III holds for many subsamples. In brief, we find similar results when (i) repeating our tests by industry or by firm or analyst

characteristics, (ii) controlling for characteristics of the covered firms, (iii) focusing on analysts covering S&P500 firms, (iv) focusing on firms and analysts with nonmissing long-term forecasts, and (v) excluding the time periods with imperfect coverage by I/B/E/S.[18]

C. Trend by Industry and Alternative Data Usage

We now show that the slope of the term structure has become steeper in industries where more alternative data are available. This evidence connects the long-run trends in analysts' forecast informativeness to alternative data and is consistent with our theory.

Section IA.X of the Internet Appendix shows that the slope of the term structure has become steeper for most FF17 industries, but not to the same extent (e.g., the steepening was stronger for Retail Stores (−9.8) than for Utilities (−2.5)). We examine whether this heterogeneity relates to analysts' differential usage of alternative data across industries measured using the text in their reports. Specifically, we search for all reports written by U.S. analysts citing at least one data vendor from the J.P.Morgan directory of alternative data vendors (Kolanovic and Smith (2019)) and construct three measures of alternative data usage; *Alternative Data Usage 1 (Alternative Data Usage 2, Alternative Data Usage 3*) corresponding to the median percentage of reports (analysts, firms) by industry-year referring to (using, covered by) alternative data over the second half of our sample period (i.e., 2000 to 2017).[19]

In Table IV, we regress the industry-level estimates for the time trend in the slope of the term structure on a constant (column (1)) and the three measures of alternative data usage (columns (2) to (4)). Column (1) shows that the average time trend across industries measured by the constant is −4.2, in line with the −4.5 reported in column (3) of Table III.[20] Columns (2) to (4) show that the three measures of alternative data usage negatively predict the trend that we observe by industry, and thus are positively associated with the steepening of the term structure. Moreover, they explain a substantial fraction of the heterogeneity of this steepening across industries, with regression R^2's ranging between 15% and 28% (and statistically insignificant constants). The rest of the table shows similar results when we consider Fama-French 48 industries.

In sum, the results in this section indicate that the informativeness of analysts' short-term forecasts has improved over time while that of their long-term forecasts has declined, resulting in a steepening of the slope of the term

[18] In Sections IA.XII and IA.XIII, we also show that the results are robust to controlling for the number of observations used to estimate equation (14) and to using interpolated forecasts to compute R^2.

[19] We describe the detailed procedure to compute each measure and report summary statistics in Sections IA.XVIII and IA.XIX of the Internet Appendix.

[20] There are only 16 instead of 17 industries in the regression because we require at least 10 years of data by industry to estimate the trend in the annual slope of the term structure.

Table IV

Trend by Industry and Alternative Data usage

This table reports estimates of the sensitivity of industry-level time trends in the slope of the term structure to the usage of alternative data in that industry. The dependent variable is the industry-level time trend in the slope of term structure, estimated by regressing the slope of the term structure (estimated by industry-year) on *Year Trend* by industry using the same specification as in column (1) of Table III. All industry-level time trend estimates are reported in the Internet Appendix (see column (6) of Table IA.VI). Each estimate measures the cumulative change in slope over the 1990 to 2017 period in the industry. A negative (positive) estimate indicates that the slope has become more (less) steep. *Alternative Data Usage* are text-based measures of usage of alternative data by analysts obtained by searching for cites of alternative data vendors in company reports of U.S. analysts (see Section IA.XVIII in the Internet Appendix for a detailed description of the procedure we use to compute each measure). In column (2), the variable is the median percentage of reports by industry-year citing at least one alternative data vendor over 2000 to 2017. In column (3), the variable is the median percentage of analysts by industry-year citing one or more alternative data vendor over 2000 to 2017. In column (4), the variable is the median percentage of firms by industry-year covered by at least one alternative data vendor over 2000 to 2017. t-Statistics in parentheses are based on robust standard errors. All specifications are estimated by WLS, with the number of observations used to estimate each industry-level time trend as weights. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Estimated Year Trend in Slope by FF17				Estimated Year Trend in Slope by FF48			
WLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Alternative Data Usage 1		−15.9*				−10.4***		
		(−1.90)				(−2.84)		
Alternative Data Usage 2			−1.2*				−0.9**	
			(−2.08)				(−2.28)	
Alternative Data Usage 3				−0.6**				−0.3**
				(−2.18)				(−2.71)
Constant	−4.2**	−0.5	1.5	2.3	−3.1**	−0.4	0.8	0.3
	(−2.88)	(−0.16)	(0.48)	(0.61)	(−2.40)	(−0.21)	(0.32)	(0.15)
N	16	16	16	16	41	41	41	41
R^2	–	14.8%	28.1%	18.6%	–	7.1%	6.9%	6.3%

structure.[21] This shift coincides with the rise of various sources of alternative data, and is more pronounced in industries in which alternative data are more frequently used by analysts. Insofar as the availability of alternative data increases the marginal benefit of obtaining short-term information (our hypothesis), this aggregate evolution and its cross-sectional variation by industry are consistent with Corollary 1. We acknowledge, however, that omitted factors could explain the long-run evolution of forecasts' informativeness and its cross-sectional variation. To address this issue, we next use the introduction of the social media platform StockTwits to capture variation in analysts' exposure to short-term-oriented data and test whether greater exposure is associated with changes in the informativeness of their forecasts.

V. The Effect of Social Media Data

A. *StockTwits Data*

StockTwits was founded in 2008 as a networking platform for investors to share their opinions about firms. 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 firms. They use $cashtags with firms' 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 2009 and 2017. For each message, we observe the user identifier, 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, as well as to firms' listing venue and "watchlist," that is, the number of users who explicitly follow the firm. We keep messages about firms that trade on NASDAQ, NYSE, NYSEArca, NYSEMkt, or that trade over-the-counter (OTC), that are present in the 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.

Figure 6 shows the evolution of the number of users and their posting intensity. The upper left panel indicates that the number of daily messages has increased from less than 1,000 in 2009 to more than 80,000 in 2017. 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 and reveal substantial and increasing heterogeneity in activity

[21] Da and Warachka (2011) find that the disparity between analysts' long-term and short-term growth forecasts in a given month predicts subsequent stock returns. Their explanation is that inattentive investors fail to account for the fact that analysts' revision of their long-run growth forecasts is less timely than the revision of their short-term growth forecasts. If the drop in analysts' long-term forecast informativeness is associated with even more sluggish adjustments in their long-term forecasts, then the predictive power of the disparity measure of Da and Warachka (2011) may have increased over time (to the extent that investors' inattention has not declined).

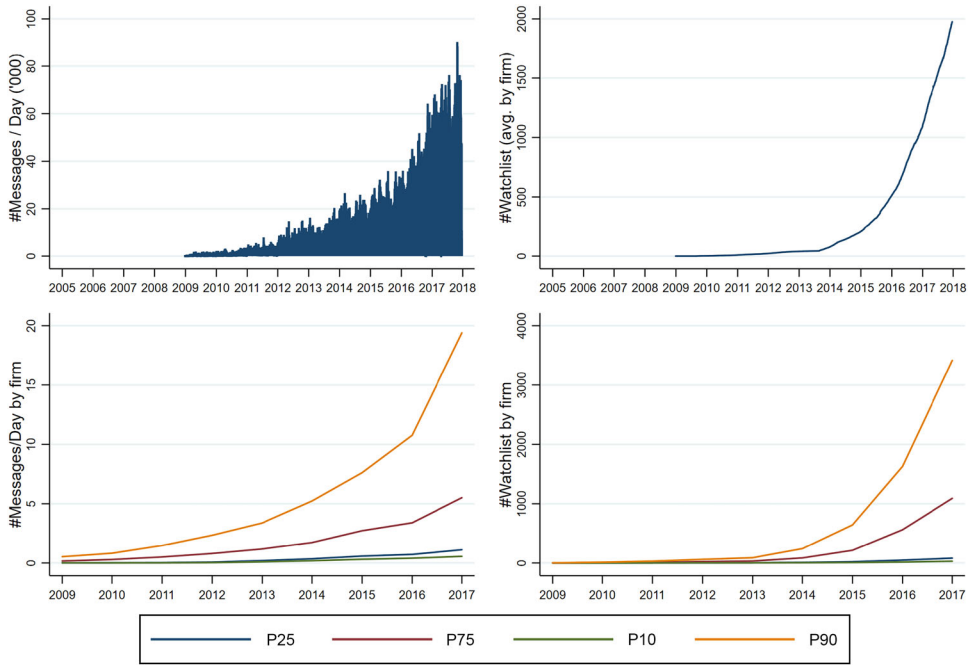

Figure 6. StockTwits' expansion. This figure shows descriptive statistics on the evolution of StockTwits between 2005 and 2017. The upper-left panel presents the total number of messages per day. The upper-right panel presents the number of users who have a given firm in their watchlist (averaged across 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 bottom-left and bottom-right panels present different percentiles of the number of messages per day per firm and the number of users who have a given firm in their watchlist, respectively. (Color figure can be viewed at wileyonlinelibrary.com)

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 tests rely on the idea that the introduction of StockTwits generates variations in analysts' exposure to short-term-oriented data because StockTwits coverage is heterogeneous (Figure 6) and analysts follow different sets of firms. This approach is relevant if the information provided by StockTwits is both short-term-oriented and used by analysts. In this section, we argue that these conditions are likely satisfied.

First, prior academic evidence shows that social media data contain incremental information about firms' future returns or earnings (see Table IA.I of Section IA.III in the Internet Appendix). We confirm that StockTwits is also short-term-oriented using "Bullish" and "Bearish" ratings issued by

StockTwits' users.[22] 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:

$$g_{j,y+h} = b_0 + b_1 Rating_{j,y} + b_2 g_{j,y-1} + \epsilon_{j,y}, \tag{15}$$

where j indexes all firms from the same fiscal year y and size quintile, $Rating_{j,y}$ 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 $y+h$.[23] We consider growth of sales, EBITDA, EBIT, or Net Income, and include $g_{j,y-1}$ to control for the effect of existing information. The coefficient b_1 measures whether current ratings predict firm growth in year $y+h$. Figure 7 displays the average of b_1 (across years and size quintiles) by horizon, and confirms that better ratings only predict firm growth only at short horizons.[24]

The second condition is that analysts use data from social media, including StockTwits. Section IA.XVII of the Internet Appendix provides an example of J.P.Morgan analysts using social media data. More systematic evidence is documented by Chi, Hwang, and Zheng (2021), who find that across different types of alternative data used by securities analysts, social media data come the second most frequently used, after app usage and on par with point-of-sale data. Among the social media data providers, StockTwits is commonly referred to as a major one, especially for discussions about firms.[25] StockTwits data feed has also been gradually integrated into all major financial information aggregation platforms used by practitioners (e.g., Bloomberg.com and Reuters.com), suggesting that analysts are commonly exposed to these data.

We provide two additional pieces of evidence on this point. 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

[22] Anecdotal evidence from industry reports also highlights the short-term nature of social media data. For example, a brochure from Deutsche Bank emphasizes the usefulness of "Estimize" (a social media platform 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 data set. 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).

[23] Note that $Rating_{j,y}$ is not a text-based measure of sentiment. Unless missing, the "Bearish" or "Bullish" rating is publicly and directly observable without ambiguity. We only consider firm-years with at least 10 messages containing a rating in the past 12 months.

[24] Year-on-year growth for the current fiscal year ($g_{j,y+0}$) is known *after* the fiscal year is over, that is, *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 run the same analysis with quarterly data. Doing so yields similar results.

[25] In their *2019 Alternative Data Handbook*, Kolanovic and Smith (2019) 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*."

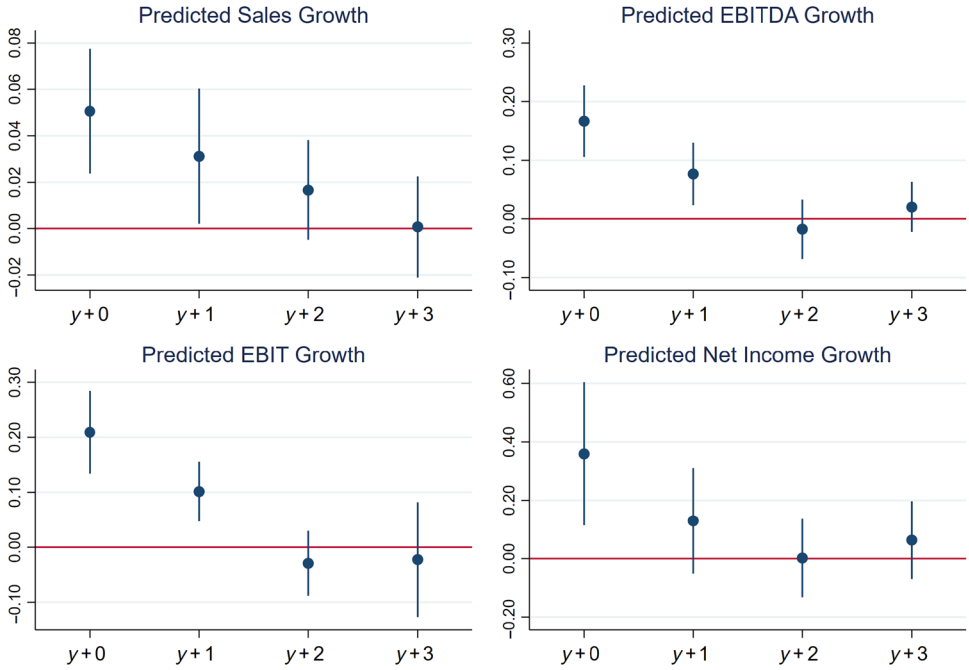

Figure 7. StockTwits ratings and firms' growth predictability. This figure shows the predictive power of "Bullish" and "Bearish" ratings issued by StockTwits users about firms' growth, by horizon. It relies on cross-sectional forecasting regressions with at least 20 observations, estimated on every fiscal year starting from 2010 by quintile of total assets, and specified as: $g_{j,y+h} = b_0 + b_1 Rating_{j,y} + b_2 g_{j,y-1} + \epsilon_{j,y}$, where j indexes all firms from the same quintile and year. $Rating_{j,y}$ is the difference between the fraction of "Bullish" and "Bearish" messages about firm j during fiscal year y provided there are least 10 messages with a nonmissing rating, and $g_{j,y+h}$ is the (year-on-year) growth 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 average of b_1 (weighted by the inverse of its standard error) across all quintiles and years y (with the associated 90% confidence interval based on standard errors clustered by year), by horizon $y+h$ (displayed on the x-axis), when g is the growth of sales (upper-left panel), EBITDA (upper-right panel), EBIT (bottom-left panel), or Net Income (bottom-right panel). (Color figure can be viewed at wileyonlinelibrary.com)

arrival from traditional data sources (see Table IA.II). Moreover, we show that analysts are more likely to upgrade (downgrade) their recommendation for a firm when more users are "Bullish" ("Bearish") about it (see Table IA.III). Second, using biographic information (analysts' last names and the first letter of their first names) from I/B/E/S over the 2000 to 2017 period, we find that 35% of (7,655 distinct) analysts' names exactly match those of StockTwits' account holders.[26]

[26] This finding is not evidence of analysts being active users. However, the mechanism that 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 in-

C. Exposure to StockTwits' Data and Forecasts' Informativeness

Our test exploits the staggered and heterogeneous expansion of StockTwits' coverage across firms to capture analysts' differential "exposure" to social media data. To identify the effect of StockTwits data, we need to isolate the variation in analysts' exposure to data that is specifically generated on StockTwits and that would not be available without this social media platform. To do so, we rely on two measures.

First, we use the number of users who have a given firm on their "watchlist" on a given day (#*Watchlist*). A user's watchlist is a list of firms she follows. StockTwits aggregates this information for each firm and reports the number of users who have the firm on their watchlist. 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.

Second, we rely on the volume of messages posted about firms. Because the number of actual messages may correlate with the arrival of information from traditional data, we estimate *hypothetical* messages. For each day, we calculate the share of total messages posted on StockTwits about each firm and compute the average share by firm, reflecting the usual daily share of messages captured by that firm. We then multiply this average share by the total number of messages posted on StockTwits on a given day to obtain the number of messages for a firm that one would expect on an "average" day. Finally, we sum the total number of such hypothetical messages for each firm in the last 30 days to obtain #*Hypothetical Messages*. By construction, the main source of variation in #*Hypothetical Messages* for a given firm is the *aggregate* number of actual messages, which should be unrelated to the regular flow of firm-level information.

Tables IA.IV and IA.V confirm that changes in firms' #*Watchlist* and #*Hypothetical Messages* are indeed uncorrelated with the arrival of information from traditional data sources. For these tests, 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 IA.IV shows no significant relationship between the number of news events and either #*Watchlist* or #*Hypothetical Messages*. Table IA.V shows similar results when using the market response to news arrival.

We next measure the exposure of a given analyst i to data specifically generated on StockTwits at time t by taking the average value of #*Watchlist* or #*Hypothetical Messages* across the firms she covers. Specifically, we define her

formation 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.

exposure as

$$\textit{Data Exposure}_{i,t} = \sum_{f=1}^{F} w_{i,f,t} K_{f,t}, \tag{16}$$

where $K_{f,t}$ is either #*Watchlist* or #*Hypothetical Messages* for firm f at time t, and $w_{i,f,t}$ is the weight of firm f (among all possibly covered firms F) in analyst i's portfolio at time t. Because we average $K_{f,t}$ across the firms followed by analyst i at time t, $w_{i,f,t}$ equals zero if she does not cover firm f at time t, and equals $1/N_{i,t}$ if she does cover it, where $N_{i,t}$ is the total number of firms she covers at t.

Armed with these measures, we estimate the effect of analysts' exposure to StockTwits data on the informativeness of their forecasts for different horizons using the specification:

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

where η_t and η_i are time (i.e., date) and analyst fixed effects, which control for common factors affecting the forecast informativeness of all analysts and for heterogeneous but time-invariant analyst-specific factors, respectively. We further control for several characteristics of the firms covered by analyst i, namely, firms' size, age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded. We average each variable across the firms covered by analyst i.[27] The sample period begins on January 1, 2005—almost five years before the first message we observe on the platform on July 13, 2009—and ends on December 31, 2017. We estimate equation (17) for each horizon subsample.

The coefficient λ measures how variation in analysts' exposure to data generated on StockTwits affects the informativeness of their forecasts at different horizons. By design, *Data Exposure* varies (within-analyst over time) due to the staggered and heterogeneous expansion of StockTwits across the firms covered by each analyst. Thus, the coefficient λ compares how the informativeness of forecasts for a given horizon changes over time for analysts with high exposure to StockTwits data relative to analysts with lower exposure. Because data generated on StockTwits are short-term-oriented, we posit that higher exposure to that data corresponds to a decrease in analysts' marginal cost to improve the precision of their short-term signal (i.e., a decrease in a in the model). According to Corollary 1, greater exposure should lead to an improvement in the informativeness of short-term forecasts (i.e., $\lambda > 0$ for small h) but possibly to a deterioration in the informativeness of long-term forecasts (i.e., $\lambda < 0$ for large h).

[27] All explanatory variables in equation (17) are winsorized at the 1% and 99% levels by date, t (unless they are log-transformed variables), are measured at $t-1$, and are defined in Appendix B.

Table V
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 to 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. *#Watchlist* is the average number of users who 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 30 days) that an analyst covers. Both variables are set to zero prior to StockTwits' introduction in 2009. *Coverage* is the number of firms that the analyst covers. *Auto* is the average earnings autocorrelation (measured by the R^2 of a regression of quarterly earnings on their lagged value) across the firms covered by an analyst. The other variables are control variables used in the analysis. Detailed variable definitions are provided in Appendix B.

	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
#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
Coverage	31,623,819	10.35	5.40	3.00	6.00	9.00	13.00	29.00
Auto	29,364,951	0.56	0.20	0.01	0.42	0.57	0.70	0.98
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
Institutional Ownership	31,620,325	0.70	0.15	0.14	0.62	0.73	0.81	0.98
Option	31,623,814	1.00	0.01	0.00	1.00	1.00	1.00	1.00
Sales growth	29,370,862	0.29	1.10	−0.43	0.05	0.12	0.26	34.93
Share turnover	31,623,814	0.01	0.01	0.00	0.00	0.01	0.01	0.14
Tech. firms	29,392,961	0.37	0.43	0.00	0.00	0.00	0.88	1.00
Tobin's Q	29,366,671	2.29	1.05	0.71	1.54	2.00	2.74	7.34

Table V presents summary statistics for the variables used in the estimation of equation (17). The sample contains 31,623,819 daily observations over the 2005 to 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 firms is typically followed by 321 users on StockTwits, and there are on average (138/30 =) 4.6 hypothetical messages about each firm, daily.

D. Results

D.1. Forecast Informativeness by Horizon: OLS

Panel A of Table VI presents OLS estimates of equation (17) by horizon subsamples, with and without control variables. To ease economic interpretation, we normalize all explanatory 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 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 0.54 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"—the value of h at which the effect of greater exposure to StockTwits on R^2 changes sign—is between one and two years.

D.2. Forecast Informativeness by Horizon: Two-Stages Least Squares

By construction (see equation (16)), variation in *Data Exposure* stems from (i) variation in data available on StockTwits about each firm (i.e., $K_{f,t}$ measured by #*Watchlist* and #*Hypothetical Messages*), and (ii) variation in coverage (i.e., $w_{i,f,t}$). As explained in Section VI.C, $K_{f,t}$ is unrelated to the arrival of information from sources other than StockTwits. However, the decision to cover a firm ($w_{i,f,t}$) might be correlated with StockTwits' expansion. If so, our estimates of the effects of analysts' exposure to StockTwits on the informativeness of their forecasts may reflect the effects of StockTwits on their coverage decision rather than on the amount of short-term-oriented data available to them.

To address this concern, we instrument *Data Exposure* with

$$\textit{Hypothetical Data Exposure}_{i,t} = \sum_{f=1}^{F} w_{i,f,2009} K_{f,t}, \tag{18}$$

where $w_{i,f,2009} = 1/N_{i,2009}$ if analyst i covers firm f in 2009 (i.e., when StockTwits started) and zero otherwise ($N_{i,2009}$ is the total number of firms covered by analyst i in 2009). Importantly, $w_{i,f,2009}$ is constant over time because we fix the set of firms we use to calculate $w_{i,f,2009}$ for each analyst i.[28] Therefore, variation in *Hypothetical Data Exposure* within analyst only reflects variation in data exposure due to StockTwits' expansion for the *same* covered firms.

Table VII presents the results from estimating equation (17) by two-stage least squares (and by horizon subsample). The coefficients on the first stage

[28] If a firm disappears from the sample, we retain the last observation of #*Watchlist* and #*Hypothetical Messages* to construct analysts' hypothetical data exposure. We obtain similar results if we drop these firms and take averages across the remaining firms.

Table VI

Data Exposure and Forecast Informativeness by Horizon (OLS)

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) at different horizons to analysts' exposure to data generated on StockTwits (equation (17)). The sample includes all available analyst-day-horizon observations between 2005 and 2017, which we split by forecasting horizon subsample. 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* captures analysts' exposure to data generated on StockTwits as defined in equation (16) and is normalized by its in-sample standard deviation. In Panel A, *Data Exposure* is based on the number of users who 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 in the past 30 days. Control variables include firms' log of total assets, log of age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Forecasts 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***	0.56***	0.40	0.23	−0.66***	−0.93***	−1.51***	−1.58***
	(3.89)	(4.13)	(1.06)	(0.63)	(−3.24)	(−4.78)	(−3.49)	(−3.27)
Analysts FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	14,055,963	13,026,178	11,489,986	10,596,175	3,916,280	3,634,874	1,496,954	1,434,373
Panel B: Proxy for Data Exposure = #Hypothetical Messages								
Data Exposure	0.66***	0.65***	0.56	0.28	−0.60*	−1.02***	−1.84***	−1.92***
	(4.39)	(4.55)	(1.27)	(0.65)	(−1.63)	(−3.44)	(−3.95)	(−3.40)
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	13,026,178	11,489,986	10,596,175	3,916,280	3,634,874	1,496,954	1,434,373

Table VII

Data Exposure and Forecast Informativeness by Horizon (2SLS)

This table presents 2SLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) at different horizons to analysts' exposure to data generated on StockTwits (equation (17)). The sample includes all available analyst-day-horizon observations between 2005 and 2017, which we split by forecasting horizon subsample. We pool horizons between three and five years because we have few observations at long horizons. The (second stage) 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 Exp.* captures analysts' exposure to data generated on StockTwits as defined in equation (16) and is normalized by its in-sample standard deviation. Hypothetical *Data Exp.* captures analysts' exposure to data generated on StockTwits if the portfolio of covered firms were fixed as defined in equation (18), and is normalized by its in-sample standard deviation. In Panel A, both *Data Exp.* and *Hypothetical Data Exp.* are based on the number of users who have the firms covered by the analyst in their watchlist. In Panel B, both *Data Exp.* and *Hypothetical Data Exp.* are based on the number of hypothetical messages posted about the firms covered by the analyst in the last 30 days. Control variables include firms' log of total assets, log of age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$h > 3$	
2SLS Dep. Variable:	First Stage *Data Exp.* (1)	Second Stage R^2 (2)	First Stage *Data Exp.* (3)	Second Stage R^2 (4)	First Stage *Data Exp.* (5)	Second Stage R^2 (6)	First Stage *Data Exp.* (7)	Second Stage R^2 (8)
Panel A: Proxy for Data Exposure = #Watchlist								
Hypothetical Data Exp.	0.57***		0.56***		0.55***		0.57***	
	(11.24)		(21.38)		(43.51)		(56.65)	
Data Exp.		0.65***		0.10		−1.07***		−1.71***
		(2.78)		(0.19)		(−3.29)		(−2.77)
Analysts FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	9,038,842	9,038,842	7,608,139	7,608,139	2,659,190	2,659,190	1,112,653	1,112,653

(*Continued*)

Table VII—*Continued*

Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$h > 3$	
2SLS Dep. Variable:	First Stage *Data Exp.* (1)	Second Stage R^2 (2)	First Stage *Data Exp.* (3)	Second Stage R^2 (4)	First Stage *Data Exp.* (5)	Second Stage R^2 (6)	First Stage *Data Exp.* (7)	Second Stage R^2 (8)
Panel B: Proxy for Data Exposure = #Hypothetical Messages								
Hypothetical Data Exp.	0.45***		0.46***		0.49***		0.40***	
	(13.25)		(14.61)		(15.51)		(10.84)	
Data Exp.		0.71**		0.59		−0.98*		−2.67*
		(1.94)		(0.76)		(−1.77)		(−1.71)
Analyst FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	9,038,842	9,038,842	7,608,139	7,608,139	2,659,190	2,659,190	1,112,653	1,112,653

are all positive and highly statistically significant because (i) analysts' coverage is persistent, and (ii) hypothetical and actual data exposure are both equal to zero before the introduction of StockTwits in July 2009. The second-stage results confirm our conclusions. An increase in analysts' *instrumented* exposure to StockTwits data leads to an improvement in the informativeness of their short-term forecasts ($h \leq 1$) and a deterioration in the informativeness of their long-term forecasts ($2 < h \leq 3$ or $h \geq 3$). Across all horizons and measures, the estimated second-stage coefficients exhibit magnitudes that are similar to those reported in Table VI, indicating that our results are not materially driven by changes in analysts' coverage.[29]

D.3. The Slope of the Term Structure

We now study whether exposure to StockTwits data has a negative effect on the slope of the term structure of forecasts' informativeness (as predicted by the last part of Corollary 1). To do so, we pool all subsample observations and modify equation (17) to allow for an interaction term between *Data Exposure* and horizon h, which we recenter at 1 and denote by h^* (i.e., $h^* = h - 1$). Specifically, we estimate:

$$R^2_{i,t,h} = \lambda_0 h^* + \lambda_1(\textit{Data Exposure}_{i,t-1}) + \lambda_2(h^* \times \textit{Data Exposure}_{i,t-1}) + \cdots + \omega_{i,t,h}. \tag{19}$$

In equation (19), λ_0 measures the (unconditional) slope of the term structure, and λ_2 the extent to which it changes with greater exposure to StockTwits data, controlling for uniform changes in $R^2_{i,t,h}$ for all h (captured by λ_1). Centering h at one is neutral on estimates for λ_2 (and λ_0), but it allows λ_1 to be interpreted as the effect of *Data Exposure* on R^2 at the one-year horizon (and not zero), and thus to detect whether the term structure simply rotates ($\lambda_2 \neq 0$ and $\lambda_1 = 0$), or also shifts either upward ($\lambda_1 > 0$) or downward ($\lambda_1 < 0$). We recenter at one because Table VI suggests that the inflection horizon is between one and two years. Corollary 1 predicts that $\lambda_2 < 0$.

We report OLS estimates of equation (19) in Panel A of Table VIII. In column (1), both λ_0 and λ_2 are negative and significant. The term structure (for the average analyst) is downward-sloping ($\lambda_0 < 0$), and greater exposure to StockTwits makes it steeper ($\lambda_2 < 0$), as predicted. Interestingly, λ_1 is not statistically different from zero, meaning that the slope changes, but informativeness at the one-year horizon does not. The term structure thus rotates around the one-year horizon but does not shift upward or downward. This finding provides evidence of a "pure" reallocation effect of exposure to StockTwits 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)).

[29] In the same vein, Table IA.XI confirms that our results hold when we focus specifically on the subset of analysts exhibiting stable coverage.

Table VIII

Data Exposure and the Slope of the Term Structure

This table presents OLS (Panel A) and 2SLS (Panel B) estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to data generated on StockTwits (equation (19)). 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* captures analysts' exposure to data generated on StockTwits as defined in equation (16) and is normalized by its standard deviation. *Hypothetical Data Exposure* captures analysts' exposure to data generated on StockTwits if the portfolio of covered firms were fixed as defined in equation (18), and is normalized by its standard deviation. *Data Exposure* and *Hypothetical Data Exposure* are based on the average number of users who have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms fin the past 30 days. 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 ($h^* = h - 1$). In Panel B, we use *Hypothetical Data Exposure* and $h^* \times$ *Hypothetical Data Exposure* as instruments. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. Control variables include firms' log of total assets, log of age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

	Forecasts Informativeness (R^2)					
Dep. Variable:	*#Watchlist*			*#Hypothetical Messages*		
Data Exposure Proxy:	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: OLS						
$h^* \times$ *Data Exposure*	−0.86***	−0.78***	−0.92***	−0.69***	−0.94***	−0.99***
	(−2.59)	(−3.06)	(−3.73)	(−2.75)	(−4.54)	(−4.72)
Data Exposure	0.13	−0.17	−0.28	0.34	−0.14	−0.2
	(0.50)	(−0.64)	(−1.09)	(1.42)	(−0.57)	(−0.81)
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
E	30,959,281	30,105,556	27,845,302	30,959,281	30,105,556	27,845,302

(*Continued*)

Table VIII—*Continued*

	Forecasts Informativeness (R^2)					
Dep. Variable:	*#Watchlist*			*#Hypothetical Messages*		
Data Exposure Proxy:	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: 2SLS (Second Stage)						
$h^* \times$ *Data Exposure*	−1.10***	−1.15***	−1.31***	−0.63	−1.16***	−1.18***
	(−2.63)	(−2.38)	(−2.90)	(−1.55)	(−2.97)	(−2.84)
Data Exposure	−0.07	−0.29	−0.28	0.54	−0.32	−0.17
	(−0.22)	(−0.71)	(−0.64)	(1.19)	(−0.68)	(−0.30)
h^*	−16.55***			−16.55***		
	(−28.44)			(−26.62)		
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	22,037,772	21,184,384	19,576,821	22,037,772	21,184,384	19,576,821

Interacting h^* with the analyst fixed effects allows us to control for permanent differences in the slope of the term structure across analysts. Likewise, interacting h^* with the date fixed effects allows us to control for the aggregate variations in the slope of the term structure that is unrelated to StockTwits' expansion.[30] Panel B of Table VIII presents 2SLS estimates of equation (19) and confirms our conclusions. The findings reported in Table VIII hold across several robustness tests, reported in Section IA.XIV of the Internet Appendix. In brief, results are similar when we control for trading volume, and thus for the potential effect of news (public or private) that is material enough to generate trading. The results are also robust to restricting our tests to analysts and firms with nonmissing long-term forecasts.

D.4. Economic Magnitude

The estimate for λ_0 in the first column of Table VIII, Panel A, implies that R^2 decreases by 16.66 percentage points for every one-year increase in the horizon. This estimate differs from the estimate reported in Section III.C because the sample period is more recent and the term structure has become steeper over time. The 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 the horizon.

The economic magnitude of this change in slope should be evaluated against normal variations. Figure 5 displays yearly estimates at the aggregate level. The standard deviation of this time series is only 1.9 over the 2005 to 2015 period, but it is 4.5 when we consider the entire 1983 to 2015 period to obtain a more precise estimate. It is 5.5 when we estimate the slope by industry and year from 2005 to 2015, and 11.8 when we estimate the slope 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%, (0.86/5.5 =) 15.6%, and (0.86/11.8 =) 7.3% of the slope long-run standard deviation at the aggregate, industry, and analyst level, respectively.[31]

[30] Controlling for permanent differences in the term structure across analysts can be achieved in two ways. If the term structure is linear, one can interact the horizon h^* with the analyst fixed effects. Because the term structure does not appear linear, we instead discretize the horizon h^* to generate annual horizon fixed effects (one-year, two-year, three-year, four-year, or five-year horizon) and interact these with analyst fixed effects (see Table VI, columns 2, 3, 5, and 6). We do not discretize h^* by day (as we do when interacting horizon with date fixed effects) because doing so creates too many categories that absorb all the variation. This full interaction approach is also used in papers that study a slope coefficient with fixed effects (see, for instance, equation (8) in Edmans, Jayaraman, and Schneemeier (2017) and the discussion that follows).

[31] This economic magnitude is larger for analysts whose names match those of a StockTwits' user account. For this subsample, and using the same specification as in column (1) of Table VIII, we find $\lambda_2 = -1.44$ (t-statistic $= -2.93$), that is, between 12.2% and 32% of the slope standard deviation.

E. Additional Predictions and Results

Corollary 1 suggests that the effects of analysts' exposure to StockTwits data should vary across analysts. Indeed, for ρ high enough or c small enough, the effect of a decrease in the cost of short-term information (a) on long-term forecast informativeness switches from negative to positive. Moreover, we show numerically in Section IA.XV of the Internet Appendix that when firms' earnings are more autocorrelated (ρ is higher) or when the cost of multitasking (c) is smaller, the negative effects of short-term-oriented data on the informativeness of long-term forecasts, and therefore the slope of the term structure of forecasts' informativeness, is smaller. In this section, we test these additional predictions. For brevity, we just consider the effects of proxies for c and ρ on the slope of the term structure (λ_2 in equation (19)) and present results regarding the effects on the level of long-term forecasts' informativeness in the Internet Appendix (Section IA.XV). Both approaches confirm the additional predictions.

First, we assess whether λ_2 is indeed more negative for analysts facing higher costs of multitasking. This should be the case for those who cover more firms because the total number of forecasting tasks (within and across firms) increases with coverage.[32] We thus count the number of firms covered by analysts and interact this variable (*Coverage*) with all variables in equation (19). Results are in Table IX. The coefficients on the triple interaction term are consistently negative, indicating that the steepening of the term structure is stronger for analysts with a higher cost of multitasking. Two other coefficients are consistently negative and highly significant. The first is the coefficient on *Coverage*, indicating that, all else equal, a greater multitasking cost negatively affects forecast informativeness for all h (which is also consistent with the model). The second is the coefficient on $h^* \times Coverage$, which shows that the steepening of the term structure increases with the cost of multitasking.

We also examine whether λ_2 is less negative when analysts cover firms whose long- and short-term earnings are more correlated. We measure the autocorrelation in a firm's earnings (ρ in the model) using the R^2 of a regression of its quarterly earnings on their lagged value (without constant) using a rolling window of two years (and requiring at least four observations). We then average these R^2 s across all firms covered by the analyst, and interact this variable (*Auto*) with all variables in equation (19). Table X shows that the coefficients on the triple interaction term are all significantly positive. Thus, as predicted, the steepening of the term structure is weaker for analysts covering firms whose earnings are *more* autocorrelated. The other coefficients are again broadly consistent with our predictions. For example, the coefficient on $h^* \times Auto$ is consistently positive and significant in four of six specifications, implying that the slope of the term structure is usually less steep for analysts covering firms with greater earnings autocorrelation.

[32] For instance, Harford et al. (2019) state 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*" (p. 2182) and show that this is the case.

Table IX

Differential Effects by Analysts' Multitasking Cost

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to 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* captures analysts' exposure to data generated on StockTwits as defined in equation (16) and is normalized by its standard deviation. *Data Exposure* is based on the average number of users who have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms in the past 30 days. 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 ($h^* = h - 1$). *Coverage* is the number of firms that the analyst covers. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. Control variables include firms' log of total assets, log of age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded, calculated using the last available financials and averaged by analyst at time $t - 1$. Detailed variable definitions are provided in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Forecasts Informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^* \times$ *Data Exposure* × *Coverage*	−0.14***	−0.06***	−0.06***	−0.10***	−0.04*	−0.06***
	(−5.71)	(−3.39)	(−3.86)	(−6.18)	(−1.64)	(−2.56)
$h^* \times$ *Data Exposure*	0.69	−0.04	−0.18	−0.06***	−0.03	−0.03
	(1.61)	(−0.10)	(−0.58)	(−2.58)	(−0.99)	(−1.35)
$h^* \times$ *Coverage*	−0.15***	−0.23***	−0.23***	−0.14***	−0.23***	−0.22***
	(−6.58)	(−8.67)	(−8.20)	(−5.96)	(−8.63)	(−8.00)
Data Exposure × *Coverage*	−0.09***	−0.05***	−0.05***	−0.06***	−0.03	−0.03
	(−3.34)	(−2.88)	(−2.41)	(−2.58)	(−0.99)	(−1.35)
#Firms	−0.22***	−0.23***	−0.25***	−0.23***	−0.24***	−0.25***
	(−5.97)	(−6.95)	(−7.08)	(−5.79)	(−6.99)	(−6.99)
Data Exposure	1.10***	0.42	0.23	0.98***	0.16	0.15
	(2.80)	(1.48)	(0.89)	(3.16)	(0.46)	(0.51)
h^*	−15.00***			−15.05***		
	(−23.62)			(−22.82)		
Analysts FE	Yes			Yes		
Date FE	Yes			Yes		
Analysts FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,281	30,105,556	27,845,302	30,959,281	30,105,556	27,845,302

Table X

Differential Effects by Earnings' Autocorrelation

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to 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* captures analysts' exposure to data generated on StockTwits as defined in equation (16) and is normalized by its standard deviation. *Data Exposure* is based on the average number of users who have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms in the past 30 days. 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 ($h^* = h - 1$). *Auto* is the average earnings' autocorrelation (measured by the R^2 of a regression of quarterly earnings on their lagged value) in analysts' portfolios and is normalized by its standard deviation. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. Control variables include firms' log of total assets, log of age, ratios of cash flow to assets, cash to assets, debt to assets, sales growth, institutional ownership, share turnover, and Tobin's Q, as well as dummy variables capturing whether firms are in the tech sector or have options traded, calculated using the last available financials and averaged by analyst at time $t - 1$. Detailed variable definitions are provided in Appendix B. t-Statistics in parentheses are based on standard errors clustered by forecasted fiscal period. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Forecasts Informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^* \times$ *Data Exposure* × *Auto*	1.49*** (6.46)	0.92*** (5.10)	0.85*** (4.67)	1.00*** (5.60)	0.60*** (3.13)	0.55*** (2.86)
$h^* \times$ *Data Exposure*	−5.17*** (−6.44)	−3.37*** (−6.94)	−3.25*** (−6.84)	−3.71*** (−6.17)	−2.59*** (−4.94)	−2.48*** (−4.77)
$h^* \times$ *Auto*	0.83*** (3.41)	0.75*** (4.32)	0.70*** (4.42)	0.41** (2.13)	0.00 (−0.02)	0.03 (0.21)
Data Exposure × *Auto*	1.49*** (6.46)	0.92*** (5.10)	0.85*** (4.67)	0.53*** (3.43)	0.50*** (3.65)	0.45*** (3.24)
Auto	1.85*** (7.63)	1.91*** (9.31)	1.35*** (6.42)	1.84*** (7.55)	1.92*** (9.11)	1.36*** (6.31)
Data Exposure	−2.33*** (−2.89)	−2.35*** (−3.99)	−2.30*** (−4.03)	−1.35*** (−2.59)	−1.59*** (−3.10)	−1.53*** (−2.89)
h^*	−18.14*** (−29.04)			−17.93*** (−26.36)		
Analysts FE	Yes			Yes		
Date FE	Yes			Yes		
Analysts FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	28,712,339	27,865,920	27,833,045	28,712,339	27,865,920	27,833,045

F. Alternative Explanations and Interpretations

Our findings are consistent with our predictions: greater exposure to StockTwits data increases the informativeness of analysts' short-term forecasts, but decreases that of their 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.

First, as discussed above, one concern is that our findings reflect changes in information available to analysts from sources other than StockTwits. However, Tables IA.IV and IA.V rule out this possibility. Second, our results may arise because uncertainty about earnings changes concurrent with the expansion of StockTwits, rendering forecasts less precise, and thus less informative. However, our R^2 measure is already normalized by total uncertainty and addresses this concern by design (see the discussion in Section IV.B). Third, our results might be related to variables affecting γ (i.e., analysts' horizon-specific incentives), including compensation schemes, career prospects, investors' demand, or brokers' internal organization. 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 equation (19)). Moreover, they should systematically coincide with the timing of StockTwits' expansion across firms, as well as the aggregate variation in messaging and following activity, while being completely unrelated to 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 it helps them learn about investors' demand for short- and long-term information. However, this learning channel requires sufficient overlap between the clients of the analysts' employers and the individuals active on StockTwits, so that analysts can actually learn about their 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 investors. Moreover, Table VII shows that our results hold for analysts covering a fixed set of firms, for whom incentives are likely stable. Thus, learning about demand is unlikely to explain our results.

VI. Conclusion

Digitization has increased the volume and diversity of alternative data available to analysts. Empirical studies suggest that these data are mainly short-term-oriented, and thus reduces the cost of obtaining short-term information. For this reason, the availability of such data affects the trade-off faced by analysts when they allocate effort between the tasks of obtaining and processing information relevant for forecasting firms' short-term versus long-term outcomes. We show theoretically that this effect can induce analysts to allocate

more effort to the collection of short-term information at the expense of long-term information. As a result, the informativeness of their short-term forecasts improves while that of their long-term forecasts can drop.

Our main contribution is to test this novel prediction. We find that, over the long run, the informativeness of analysts' short-term forecasts has improved while that of their long-term forecasts has declined. In line with our theory, these opposing trends are more pronounced in industries in which analysts use more alternative data. We also show that an increase in analysts' exposure to short-term-oriented data (via the introduction of the social media platform StockTwits) is associated with a drop in the informativeness of their long-term forecasts, while the informativeness of their short-term forecasts improves. This finding raises many new and interesting questions about the effects of short-term-oriented data. For instance, their availability could affect real outcomes (e.g., corporate investment) via their effects on the informativeness of long-term forecasts.[33]

Initial submission: July 20, 2021; Accepted: September 16, 2022
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix A: R^2 Estimation Procedure

This appendix shows how to estimate the informativeness of analysts' forecasts, R^2, based on the initial sample of 9,129,282 unique forecasts and realizations described in Section III.A. We illustrate our procedure with a fictitious analyst XYZ covering six firms (A, B, C, D, E, and 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, the next fiscal period, and the subsequent fiscal periods. Since the measure is horizon-specific, forecasts related 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 related 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.[34] Column (1) of Table A.I shows the last available earnings forecasts made by XYZ for A, B, C, D, E, and F as of

[33] Derrien and Kecskes (2013) show that a drop in analysts' coverage has a negative effect on corporate investment. This suggests that the information produced by analysts matter for real decisions.

[34] For example, if as of January 19, 2007, the latest earnings forecast for B made by XYZ were older than 365 days, we would compute the R^2 without firm B.

Table A.I
Example of R^2 Computation for Analyst XYZ on January 19, 2007

Forecasted Fiscal Period: December 31, 2008

Firm	Latest Forecast ($M) (1)	Realized Earnings ($M) (2)	Total Assets ($M) (3)	Earnings Report Date (4)	Latest Normalized Forecast (f_j) (5)	Realized Normalized Earnings (e_j) (6)
A	110	66	1,100	March 31, 2009	0.10	0.06
B	30	18	250	March 31, 2009	0.12	0.07
C	59	15	735	March 31, 2009	0.08	0.02
D	740	538	6,725	March 31, 2009	0.11	0.08
E	1,021	1,225	10,210	March 31, 2009	0.10	0.12
F	7	3	55	March 31, 2009	0.12	0.06

January 19, 2007. The actual realized earnings for fiscal year 2008 are in column (2).[35]

- 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 A.I, column (3). Earnings forecasts (f_j) and realized earnings (e_j) after normalization are reported in columns (5) and (6).[36]
- Step 4: Estimate equation (14) by ordinary least squares (OLS) and compute R^2. Regress e_j on f_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 regression. The R^2 is set to zero if f_j negatively predicts e_j (i.e., if $k_1 < 0$ in equation (14)). It is set to missing if there are fewer than three or more than 30 observations in the regression, or if k_1 is missing after trimming the regression coefficient at the 1% level in each tail.[37] In Table A.I, the R^2 of the regression of e_j (column (6)) on f_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

[35] Notice that 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 first estimating all unobserved forecasts between two consecutive observable forecasts by linear interpolation, and then using these interpolated forecasts instead of the last available forecast to compute R^2 daily.

[36] Our results are robust to different normalization approaches. In this example, total assets could be measured as of December 31, 2006, that is, 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).

[37] 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 equation (14) with few observations.

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) of Table A.I 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 R^2 from Step 4 of 14.9% is thus 2.20 years (802 days, divided by 365).

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 one day and five years across 14,379 distinct analysts.

Appendix B: Variable Definitions

Variable	Definition
All variables below are *analyst-level* variables	
#Firms	Number of firm observations used to estimate equation (14).
Coverage	Total number of distinct firms covered by an analyst on a given day.
h	Number of days between the date at which we observe 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 a given day 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 30 days (from $t - 30$ to $t - 1$).
#Hypothetical Messages	Number of hypothetical StockTwits' messages posted about a given firm over the last 30 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 with 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, measured by the R^2 of a regression of ibq over the lag of ibq over the last two years (without constant). We require that the regression has at least four observations.
Cash flow to assets	$(ib + dp)/at$ (from last available financial statements in Compustat).

Variable	Definition
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).
Institutional ownership	*instown pct* (from last available record in WRDS Thomson Reuters Institutional (13f) Holdings - Stock Ownership Summary).
Option	Dummy variable indicating whether the firm has option listing covered in OptionMetrics (from 1996 onward).
Sales growth	Growth in sales (*sale* item) (from last available financial statements in Compustat)
Share turnover	$vol/(shrout * 1,000)$ (from last available record in the CRSP).
Tech. firms	Dummy variable indicating whether the firm is in one of the following three-digit SIC: 283, 357, 366, 367, 382, 384, 481, 482, 489, 737, or 873 (see Kile and Phillips (2009)).
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 C: Derivations in the Model

A. *Proof of Equation (6)*

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

$$\begin{aligned} \frac{\partial \overline{W}}{\partial f_{st}} &= -2\gamma(f^*_{st} - \mathsf{E}(\theta_{st} \mid s_{st}, s_{lt})) = 0 \\ \frac{\partial \overline{W}}{\partial f_{lt}} &= -2(1-\gamma)(f^*_{lt} - \mathsf{E}(\theta_{lt} \mid s_{st}, s_{lt})) = 0. \end{aligned} \tag{C1}$$

Solving for f^*_{st} and f^*_{lt} and using the fact that s_{lt} is uninformative about θ_{st}, we obtain equation (6). It is straightforward that the second-order conditions are satisfied.

B. *Proof of Equation (7)*

Substituting equation (6) into (5), we obtain

$$\begin{aligned} \mathsf{E}(\overline{W}(f^*_{st}, f^*_{lt}; s_{st}, s_{lt})) &= \omega - \gamma\mathsf{E}((\mathsf{E}(\theta_{st}|s_{st}) - \theta_{st})^2) - (1-\gamma)\mathsf{E}((\mathsf{E}(\theta_{lt}|s_{st}, s_{lt}) - \theta_{lt})^2), \\ &= \omega - \gamma\mathsf{E}(\mathsf{var}(\theta_{st}|s_{st})) - (1-\gamma)\mathsf{E}(\mathsf{var}(\theta_{lt}|s_{lt}, s_{st})), \\ &= \omega - q(\beta,\gamma)\mathsf{var}(\theta_{st}|s_{st}) - (1-\gamma)\mathsf{var}(e_{lt}|s_{lt}). \end{aligned} \tag{C2}$$

The last line in equation (C2) 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 common component (θ_{st}) and the unique component (e_{lt}) of the long-term earnings are independent.

C. *Proof of Proposition 1*

Substituting $\mathrm{var}(\theta_{st}|s_{st})$ and $\mathrm{var}(e_{lt}|s_{st}, s_{lt})$ in the analyst's objective function in equation (9) by their expressions in equation (4), we obtain that the first-order conditions for an interior solution to the analyst's optimization problem at date 0 are

$$\begin{aligned} q(\beta,\gamma)\psi_{st}\sigma_{st}^2 - 2az_{st}^* - cz_{lt}^* = 0, \\ (1-\gamma)\psi_{lt}\sigma_e^2 - 2bz_{lt}^* - cz_{st}^* = 0. \end{aligned} \tag{C3}$$

It is then straightforward to check that the solution to this system of equations is given by (z_{st}^*, z_{lt}^*) as defined in equation (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 $0 \leq z_h \leq (\psi_h)^{-1}$ (with strict inequalities for an interior solution) and $4ab > c^2$.

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:

$$\frac{\psi_{lt}\sigma_e^2}{\psi_{st}\sigma_{st}^2} \leq \frac{2b \times q(\beta,\gamma)}{c(1-\gamma)}, \tag{C4}$$

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

$$\frac{c \times q(\beta,\gamma)}{2a(1-\gamma)} \leq \frac{\psi_{lt}\sigma_e^2}{\psi_{st}\sigma_{st}^2}. \tag{C5}$$

It is immediate that if conditions (C4) and (C5) are satisfied, then the condition $4ab > c^2$ is satisfied as well. Finally, it is easily checked that these two conditions are equivalent to

$$c < \bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt}), \tag{C6}$$

where

$$\bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt}) = \mathrm{Min}\left\{ \frac{2\frac{\psi_{lt}\sigma_e^2}{\psi_{st}\sigma_{st}^2}a(1-\gamma)}{q(\beta,\gamma)}, \frac{2bq(\beta,\gamma)}{\frac{\psi_{lt}\sigma_e^2}{\psi_{st}\sigma_{st}^2}(1-\gamma)} \right\}.$$

Moreover, under the condition $c < \bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt})$, z_h^* decreases with c for $h \in \{st, lt\}$. Thus, a sufficient condition for $z_h^* < \psi_h^{-1}$ is that this condition

is satisfied when $c = 0$, which is the case if $\psi_{st} < (2a/\sigma_{st}^2 q(\beta, \gamma))^{\frac{1}{2}}$ and $\psi_{st} < (2b/\sigma_e^2(1-\gamma))^{\frac{1}{2}}$.

Finally, using the expressions for z_{st}^* and z_{lt}^* in Proposition 1, we have that

$$\begin{aligned} \frac{\partial z_{st}^*}{\partial a} &= -\frac{4b}{(4ab - c^2)} z_{st}^* < 0, \\ \frac{\partial z_{lt}^*}{\partial a} &= \frac{2c}{(4ab - c^2)} z_{st}^* > 0 \quad \text{if} \quad c > 0. \end{aligned} \tag{C7}$$

D. Proof of Equations (12) and (13)

By definition, $\mathrm{var}(\theta_{lt} | f_{lt}^*) = \mathrm{E}((\theta_{lt} - \mathrm{E}(\theta_{lt} \mid f_{lt}^*))^2 \mid f_{lt}^*)$. Given that $f_{lt}^* = \mathrm{E}(\theta_{lt} \mid s_{st}, s_{lt})$, we deduce that: $\mathrm{var}(\theta_{lt} | f_{lt}^*) = \mathrm{E}((\theta_{lt} - \mathrm{E}(\theta_{lt} \mid s_{st}, s_{lt}))^2 \mid f_{lt}^*)$. The law of iterated expectations implies that $\mathrm{var}(\theta_{lt} | f_{lt}^*) = \mathrm{E}(\mathrm{var}(\theta_{lt} \mid s_{st}, s_{lt})) \mid f_{lt}^*)$. Since $\mathrm{var}(\theta_{lt} \mid s_{st}, s_{lt})$ 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: $\mathrm{var}(\theta_{lt} | f_{lt}^*) = \mathrm{var}(\theta_{lt} \mid s_{st}, s_{lt})$. Finally, as $\theta_{lt} = \beta\theta_{st} + e_{lt}$, we have that

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

Given that s_{lt} and s_{st} are independent and e_{lt} and θ_{st} are unconditionally independent, we have $\mathrm{cov}(\theta_{st}, e_{lt} \mid s_{st}, s_{lt}) = 0$. It follows from equation (4) that $\mathrm{var}(\theta_{lt} | f_{lt}^*) = \beta^2\sigma_{st}^2(1 - \psi_{st} z_{st}^*) + \sigma_e^2(1 - \psi_{lt} z_{lt}^*)$, that is, since $\mathrm{var}(\theta_{lt}) = \beta^2\sigma_{st}^2 + \sigma_e^2$,

$$\mathrm{var}(\theta_{lt} \,|\, f_{lt}^*) = \mathrm{var}(\theta_{lt}) - \beta^2 \psi_{st} \sigma_{st}^2 z_{st}^* - \sigma_e^2 \psi_{lt} z_{lt}^*. \tag{C8}$$

Therefore, using the definition of R_{lt}^2 and the fact that $\mathrm{var}(\theta_{lt}) = \beta^2\sigma_{st}^2 + \sigma_e^2$, we obtain

$$R_{lt}^2 = \frac{\beta^2\sigma_{st}^2}{\beta^2\sigma_{st}^2 + \sigma_e^2} \psi_{st} z_{st}^* + \frac{\sigma_e^2}{\beta^2\sigma_{st}^2 + \sigma_e^2} \psi_{lt} z_{lt}^*, \tag{C9}$$

which yields the expression for R_{lt}^2 in equation (13) given that $\rho^2 = \frac{\beta^2\sigma_{st}^2}{\beta^2\sigma_{st}^2 + \sigma_e^2}$. The derivation of the expression for R_{st}^2 follows the same step and is omitted for brevity.

E. Proof of Corollary 1

Differentiating equation (12) with respect to the marginal cost of producing short-term information, a, we obtain

$$\frac{\partial R_{st}^2}{\partial a} = \psi_{st} \left(\frac{\partial z_{st}^*}{\partial a} \right) < 0 \tag{C10}$$

since $\frac{\partial z^*_{st}}{\partial a} < 0$ (see equation (C7)). This yields the second part of the corollary. Moreover, differentiating equation (13) with respect to a and using equation (C10), we obtain

$$\frac{\partial R^2_{lt}}{\partial a} = \rho^2 \psi_{st} \frac{\partial z^*_{st}}{\partial a} + (1-\rho^2)\psi_{lt}\left(\frac{\partial z^*_{lt}}{\partial a}\right). \tag{C11}$$

Since $\frac{\partial z^*_{lt}}{\partial a} = -\frac{c}{2b}\frac{\partial z^*_{st}}{\partial a}$ (see equation (C7)), we can rewrite the previous equation as

$$\frac{\partial R^2_{lt}}{\partial a} = \frac{\partial z^*_{st}}{\partial a}\left(\rho^2 \psi_{st} - (1-\rho^2)\psi_{lt}\frac{c}{2b}\right). \tag{C12}$$

As $\frac{\partial z^*_{st}}{\partial a} < 0$, we obtain that $\frac{\partial R^2_{lt}}{\partial a} > 0$ if and only if $\rho < (\frac{c\psi_{lt}}{2b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$. Thus, a decrease in a reduces the informativeness of long-term forecasts if and only if $\rho < (\frac{c\psi_{lt}}{2b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$, as claimed in the first part of the corollary. Last, observe that

$$\Delta = R^2_{lt} - R^2_{st} = -(1-\rho^2)(\psi_{st} z^*_{st} - \psi_{lt} z^*_{lt}), \tag{C13}$$

where the second equality follows from equations (12) and (13). Thus, using the fact that $\frac{\partial z^*_{lt}}{\partial a} = -\frac{c}{2b}(\frac{\partial z^*_{st}}{\partial a})$, we have that

$$\frac{\partial \Delta}{\partial a} = -(1-\rho^2)\frac{\partial z^*_{st}}{\partial a}\left(\psi_{st} + \psi_{lt}\frac{c}{2b}\right) > 0, \tag{C14}$$

where the last inequality follows from $\frac{\partial z^*_{lt}}{\partial a} < 0$.

REFERENCES

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

Bai, Jennie, Thomas Philippon, and Alexi Savov, 2016, Have financial markets become more informative?, *Journal of Financial Economics* 122, 625–654.

Bandyopadhyay, Sati, Lawrence Brown, and Gordon Richardson, 1995, Analysts' use of earnings forecasts in predicting stock returns: Forecast horizon effects, *International Journal of Forecasting* 11, 429–445.

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

Bradshaw, Mark T., 2004, How do analysts use their earnings forecasts in generating stock recommendations?, *The Accounting Review* 79, 25–50.

Campbell, John, and Samuel Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *Review of Financial Studies* 21, 1509–1531.

Chen, Long, Zhi Da, and Xinlei Zhao, 2013, What drives stock price movements?, *Review of Financial Studies* 26, 841–876.

Chi, Feng, Byoung Hwang, and Yaping Zheng, 2021, The use and usefulness of big data in finance: Evidence from financial analysts, Working paper, Cornell University.

Cookson, J. Anthony, Joseph E. Engelberg, and William Mullins, 2020, Does partisanship shape investor beliefs? Evidence from the COVID-19 pandemic, *Review of Asset Pricing Studies* 10, 863–893.

Cookson, J. Anthony, Joseph E. Engelberg, and William Mullins, 2022, Echo chambers, *Review of Financial Studies* 36, 450–500.
Cookson, J. Anthony, and Marina Niessner, 2020, Why don't we agree? Evidence from a social network of investors, *Journal of Finance* 75, 173–228.
Copeland, Tom, Aaron Dogloff, and Alberto Moel, 2004, The role of expectations in explaining the cross-section of stock returns, *Review of Accounting Studies* 9, 149–188.
Da, Zhi, and Mitch Warachka, 2011, The disparity between long-term and short-term forecasted earnings growth, *Journal of Financial Economics* 100, 424–442.
Derrien, François, and Ambrus Kecskes, 2013, The real effects of financial shocks: Evidence from exogenous changes in analyst coverage, *Journal of Finance* 68, 1407–1440.
Dugast, Jérôme, and Thierry Foucault, 2018, Data abundance and asset price informativeness, *Journal of Financial Economics* 130, 367–391.
Dugast, Jérôme, and Thierry Foucault, 2022, Equilibrium data mining and data abundance, *Journal of Finance* (forthcoming).
Edmans, Alex, Sudarshan Jayaraman, and Jan Schneemeier, 2017, The source of information in prices and investment-price sensitivity, *Journal of Financial Economics* 126, 74–96.
Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran, 2021, Where has all the data gone?, *Review of Financial Studies* 35, 3101–3138.
Farboodi, Maryam, and Laura Veldkamp, 2020, Long-run growth of financial data technology, *American Economic Review* 110, 2485–2523.
Gao, Meng, and Jiekun Huang, 2020, Informing the market: The effect of modern information technologies on information production, *Review of Financial Studies* 33, 1367–1411.
Gerken, William C., and Marcus O. Painter, 2022, The value of differing points of view: Evidence from financial analysts' geographic diversity, *Review of Financial Studies* 36, 409–449.
Giannini, Robert, Paul Irvine, and Tao Shu, 2019, The convergence and divergence of investors' opinions around earnings news: Evidence from a social network, *Journal of Financial Markets* 42, 94–120.
Goldstein, Itay, and Liyan Yang, 2015, Information diversity and complementarities in trading and information acquisition, *Journal of Finance* 70, 1723–1765.
Green, T. Clifton, Ruoyan Huang, Quan Wen, and Dexin Zhou, 2019, Crowdsourced employer reviews and stock returns, *Journal of Financial Economics* 134, 236–251.
Grennan, Jillian, and Roni Michaely, 2020, FinTechs and the market for financial analysis, *Journal of Financial and Quantitative Analysis* 56, 1–31.
Grennan, Jillian, and Roni Michaely, 2021, Artificial intelligence and the future of work: Evidence from analysts, Working paper, Duke University.
Harford, Jarrad, Feng Jiang, Rong Wang, and Fei Xie, 2019, Analyst career concerns, effort allocation, and firms' informational environment, *Review of Financial Studies* 32, 2179–2224.
Hilary, Gilles, and Charles Hsu, 2013, Analyst forecast consistency, *Journal of Finance* 68, 271–297.
Hirshleifer, David, Yaron Levi, Ben Lourie, and Siew Hong Teoh, 2019, Decision fatigue and heuristic analyst forecasts, *Journal of Financial Economics* 133, 83–98.
Hong, Harrison, and Marcin Kacperczyk, 2010, Competition and bias, *Quarterly Journal of Economics* 125, 1683–1725.
Huang, Shiyang, Yan Xiong, and Liyan Yang, 2022, Skill acquisition and data sales, *Management Science* 68, 6116–6144.
Jung, Boochun, Philip Shane, and Yanhua Yang, 2012, Do analysts long-term growth forecasts matter? Evidence from stock recommendations and career outcomes, *Journal of Accounting and Economics* 53, 55–76.
Kile, Charles, and Mary Phillips, 2009, Using industry classification codes to sample high-technology firms: Analysis and recommendations, *Journal of Accounting, Auditing & Finance* 24, 35–58.
Kolanovic, Marko, and Robert Smith, 2019, *Big Data and AI Strategies, 2019 Alternative Data Handbook* (J.P. Morgan).
Martin, Ian, and Stefan Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.

Merkley, Kenneth, Roni Michaely, and Joseph Pacelli, 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, David P., and Elizabeth Plummer, 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, Roxana, 2020, Financial innovation and the inequality gap, Working paper, University of Lausanne.

Srinidhi, Bin, Sidney Leung, and Bikki Jaggi, 2009, Differential effects of Regulation FD on short- and long-term analyst forecasts, *Journal of Accounting and Public Policy* 28, 401–418.

Thesmar, David, and Tim de Silva, 2021, Noise in expectations: Evidence from analysts forecasts, Working paper, MIT.

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2022, Man versus machine learning: The term structure of earnings expectations and conditional biases, *Review of Financial Studies* 36, 2361–2396.

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

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.