Online Appendix for

Does Alternative Data Improve Financial Forecasting? The Horizon Effect

(not intended for publication)

June 15, 2022

Contents

1 Dividing Forecasting Tasks

In our model, the analyst is in charge of two forecasting tasks and bears a multi-tasking cost. One may wonder whether she would not be better off assigning these two tasks to two different agents to save on the multitasking cost. In this section, we identify two reasons why this may not be optimal: (i) Duplication of fixed cost of information production and (ii) Agency costs. In particular, in Section 1.1, we provide the condition on the parameters of the model such that dividing the tasks is suboptimal (see eq.(4)).

1.1 Duplication of fixed costs of information production

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

That is, the analyst is better off not dividing the tasks if the maximum value of the multi-tasking costs (that incurred when the agents choose the largests optimal efforts, i.e., obtained when $c = 0$) is less than the incremental cost incurred by dividing the tasks between two agents (C_0). Using the expressions for $z^*_{st}(0)$ and $z^*_{lt}(0)$ in Proposition 1, the previous condition is equivalent to:

$$c \leq \frac{4C_0}{q(\beta,\gamma)(1-\gamma)(\psi_{st}\sigma^2_{st}\psi_{lt}\sigma^2_e)}. \tag{4}$$

Thus, we obtain that if $c \leq \text{Min}\{\bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt}), \frac{4C_0}{q(\beta,\gamma)(1-\gamma)(\psi_{st}\sigma^2_{st}\psi_{lt}\sigma^2_e)}\}$ (where $\bar{c}(\beta,\gamma,a,b,\psi_{st},\psi_{lt})$ is defined in the proof of Proposition 1), Proposition 1 holds and it is not optimal for the analyst to hire two agents, despite the multi-tasking cost.

1.2 Agency costs.

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

2 Shock on ψ_{st}

In this section, we show that if $\rho < (\frac{c\psi_{lt}}{4b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$ then (i) the informativeness of the analyst's short-term forecast increases with the marginal return on effort for obtaining short-term information (ψ_{st}), i.e., $\frac{\partial R^2_{st}}{\partial \psi_{st}} > 0$ and (ii) the informativeness of the analyst's long-term forecast decreases with the marginal return on effort for obtaining short-term information ($\frac{\partial R^2_{lt}}{\partial \psi_{st}} < 0$). It is direct from Proposition 1 that

$$\frac{\partial z^*_{st}}{\partial \psi_{st}} = \frac{2bq(\beta,\gamma)\sigma^2_{st}}{4ab-c^2} > 0, \quad \text{and} \quad \frac{\partial z^*_{lt}}{\partial \psi_{st}} = -\frac{cq(\beta,\gamma)\sigma^2_{st}}{4ab-c^2} < 0. \tag{5}$$

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

Using eq.(12) in the main text, it immediately follows that $\frac{\partial R^2_{st}}{\partial \psi_{st}} = \psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}} + z^*_{st} > 0$. Moreover, using eq(13) in the main text, we obtain

$$\frac{\partial R^2_{lt}}{\partial \psi_{st}} = (1-\rho^2)\psi_{lt}\frac{\partial z^*_{lt}}{\partial \psi_{st}} + \rho^2\frac{\partial R^2_{st}}{\partial \psi_{st}}. \tag{6}$$

Using the fact that $\frac{\partial z^*_{lt}}{\partial \psi_{st}} = -(\frac{c}{2b})\frac{\partial z^*_{st}}{\partial \psi_{st}}$ and $\frac{\partial R^2_{st}}{\partial \psi_{st}} = \psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}} + z^*_{st}$, we deduce that

$$\frac{\partial R^2_{lt}}{\partial \psi_{st}} = -(1-\rho^2)\psi_{lt}(\frac{c}{2b})\frac{\partial z^*_{st}}{\partial \psi_{st}} + \rho^2(\psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}} + z^*_{st}). \tag{7}$$

Now, as $0 < z^*_{st} < \psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}}$ (see the expressions for z^*_{st} in the text and $\frac{\partial z^*_{st}}{\partial \psi_{st}}$ in eq.(5)), we deduce that

$$\frac{\partial R^2_{lt}}{\partial \psi_{st}} < -(1-\rho^2)\psi_{lt}(\frac{c}{2b})\frac{\partial z^*_{st}}{\partial \psi_{st}} + 2\rho^2\psi_{st}\frac{\partial z^*_{st}}{\partial \psi_{st}}. \tag{8}$$

Thus, as $\frac{\partial z^*_{st}}{\partial \psi_{st}} > 0$, a sufficient condition for $\frac{\partial R^2_{lt}}{\partial \psi_{st}} < 0$ is $\rho < (\frac{c\psi_{lt}}{4b\psi_{st}+c\psi_{lt}})^{\frac{1}{2}}$.

3 Literature Review on Predictability with Alternative Data

We have reviewed 26 academic papers using alternative datasets (bibliographical references are given in the last section of this internet appendix). Table A1 shows the type of alternative data used in each paper, and whether the authors find evidence of predictive power for short-term and long-term outcomes (real outcomes or stock returns). All papers (but one) show that alternative data predict fundamentals or returns in the short-term, defined as an horizon shorter than one year. One paper discusses long-term predictability, but reports insignificant results. We could not find any academic evidence that alternative datasets contain information about firms' long-term prospects.

Table A1: Existing Literature on Data Predictability

This table reviews the papers published in Finance and Accounting journals that investigate the usefulness of various recently available data for predicting firm real outcomes and stock returns. For both outcomes, we indicate whether the paper provides evidence of short-term (i.e., less than a year) and long-term (i.e., more than a year) predictability. "yes" indicates that predictability is found. "no" indicates that predictability is rejected. "NA" indicates that predictability is not tested.

Paper	Data type	Predictability for:			
		Real Outcomes		Stock Returns	
		Short-term	Long-term	Short-term	Long-term
Green, Huang, Wen, Zhou (2019)	Employer ratings	yes	no	yes	no
Katona, Painter, Patatoukas, Zeng (2021)	Satellite Images	NA	NA	yes	NA
Mukherjee, Panayotov, Shon (2021)	Weather Data	NA	NA	yes	NA
Leung, Wong, Wong (2019)	Social Media	yes	NA	yes	NA
Froot, Kang, Ozik, Sadka (2016)	Mobile devices	yes	NA	yes	NA
Umar (2022)	Social Media	yes	NA	yes	NA
Tang (2017)	Social Media	yes	NA	NA	NA
Huang (2018)	Product Reviews	yes	NA	yes	NA
Zhu (2019)	Satellite Images	yes	NA	yes	NA
Jame, Johnston, Markovm, Wolfe (2016)	Social Media	yes	NA	yes	NA
Chen, De, Hu, Hwang (2014)	Social Media	yes	NA	yes	NA
Bartov, Faurel, Mohanram (2018)	Social Media	yes	NA	yes	NA
Kelley, Tetlock (2013)	Retail Orders	yes	NA	yes	NA
Bollen, Mao, Zeng (2011)	Social Media	NA	NA	yes	NA
Da, Engelberg, Gao (2011)	Google Searches	yes	NA	yes	NA
Farrell, Green, Jame, Markov (2020)	Social Media	yes	NA	yes	NA
Hirshleifer, Shumway (2003)	Weather Data	NA	NA	yes	NA
Goetzmann, Kim, Kumar, Wang (2015)	Weather Data	NA	NA	yes	NA
Giannini, Irvine, Shu (2017)	Social Media	NA	NA	yes	NA
Choi, Varian (2009)	Google Trend	yes	NA	NA	NA
Wu, Brynjolfsson (2015)	Google Trend	yes	NA	NA	NA
Tumarkin, Whitelaw (2001)	Social Media	NA	NA	no	NA
Antweiler, Frank (2004)	Social Media	NA	NA	yes	NA
Hirschey, Richardson, Scholz (2000)	Social Media	NA	NA	yes	NA
Drake, Roulstone, Thornock (2012)	Google Searches	NA	NA	yes	NA
Gu, Kurkov (2020)	Social Media	yes	NA	yes	NA

4 EPS to Net Income Forecast Conversion

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

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

5 Why not Use Long-Term Growth Forecasts?

Analysts sometimes disclose, in addition to their earnings forecasts, a forecast about long-term growth. Specifically, a long-term growth (LTG) forecast in percent is reported instead of earnings forecasts in dollar amounts for more distant and specific fiscal periods. These LTG forecasts have been used in the literature, either directly to understand belief formation (e.g., Bordalo, Gennaioli, La Porta, and Shleifer (2019)), or indirectly to estimate the cost of capital (e.g., Chen, Da, and Xhao (2013)). LTG forecasts are however not well suited for our purpose because the horizon of these forecasts is unclear, making it difficult to assign them to actual realizations (and hence measure precisely their informativeness by horizon). After reading several reports from analysts, we indeed find substantial heterogeneity in how analysts define the horizon of their LTG forecasts (when they provide this definition). Some

refer to earnings growth for the next five years, others use the next three years. Many refer to 3-5 year growth, without any further detail. Moreover, the base year for the growth estimate also varies. It can be the last historical fiscal year, the current fiscal year, the next fiscal year, or the subsequent one. Often, this base year is undefined.

6 Forecast Informativeness with Biased Analysts

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

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

where f_{hi} is the analyst's forecast about firm i's earnings, θ_{hi}, at horizon h, Ω is the analyst's information and $\widetilde{b}_{hi}$ is the analyst's bias, which can be random. This bias is independent from $\theta_{hi} - \mathrm{E}(\theta_{hi} \,|\Omega)$ (i.e., the bias cannot be used by the analyst to forecast θ_{hi} beyond and above the information in Ω) and from $\mathrm{E}(\theta_{hi} \,|\Omega)$. That is, $cov(\tilde{b}, \theta_{hi} - \mathrm{E}(\theta_{hi} \,|\Omega)) = 0$ and $cov(\mathrm{E}(\theta_{hi} \,|\Omega)\,, \tilde{b}) = 0$. The particular case in which the bias is constant satisfied these assumptions.

In our model, $\widetilde{b}_{hi} = 0$ (the analyst is unbiased). On average, the analyst's bias at horizon h is

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

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

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

Thus, when an analyst is biased, her expected forecasting error is the sum of: (i) the precision of her unbiased forecast ($\mathrm{Var}(\theta_{hi} \,|\Omega)^{(-1)}$), (ii) the variance of her bias, or (iii) her expected bias ($\mathrm{E}(\widetilde{b}_{hi})$. We are interested in measuring (i) (i.e., the quality of the analyst's unbiased

forecast) and not (ii) and (iii) (distorsions induced by the bias in the analyst's forecasts). For instance, as pointed out by Hilary and Hsu (2013), the analyst's average squared error can be large because her expected bias is large, even though the analyst's unbiased forecast is perfect ($\text{Var}(\theta_{hi}\,|\Omega) = 0$).

As shown below, in contrast to the average analyst's squared forecasting error, our measure of analysts' forecast informativeness is not affected by the analysts' forecasting bias and identical to the informativeness of their unbiased forecasts when the bias is constant.

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

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

with $\widehat{k}_0 = (E(\theta_{hi})(1-\widehat{k}_1) - \widehat{k}_1\text{E}(\widetilde{b}_{hi}))$ and $\widehat{k}_1 = \frac{Var(f^*_{hi})}{Var(f_{hi})}$ with $\hat{k}_1 \leq 1$. Note that $\hat{k}_1 = 1$ when the analyst's bias is constant.

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

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

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

7 Analysts' Forecasting Activity and Recommendations

Our test in Section VI builds on the assumption that (some) analysts use StockTwits data as a complementary source of information (see discussion in Section VI.B). Table A2 and Table A3 report results (discussed in section VI.B) that are consistent with this assumption.

In Table A2, Column (1) shows that analysts are more likely to issue a new forecast on a given firm and day following an increase in StockTwits activity, as measured by the number of actual messages posted about the firm over the last 30 days. Column (2) shows that this result survives when controlling for trading volume, and thus for the possible effects of contemporaneous news (public or private) that is material enough to generate trading. Columns (3) and (4) show that this result continues to hold on days without news arrival from traditional data sources (which we identify using Capital IQ Key Developments), and thus mitigate the concern that news arrival (affecting both analysts' forecasts and social media activity) confounds the relationship documented in Column (1).

In Table A3, Column (1) shows that the recommendation of an analyst on a given firm and day is positively (negatively) related to the fraction of StockTwits users whose opinion is "Bullish" ("Bearish"). When more users are "Bullish" ("Bearish"), analysts are more likely to upgrade (downgrade) their recommendation. The economic magnitude of this effect is small, but it is highly significant. Columns (2) and (3) show that this result continues to hold on days without news arrival from traditional data sources (which we identify using Capital IQ Key Developments), and thus mitigate the concern that news arrival (affecting both analysts' recommendations and users' ratings) confounds the relationship documented in Column (1) of Table A3.

Table A2: Social Media Data and Analysts' Forecasting Activity

This table presents OLS estimates of the sensitivity of analysts' propensity to issue new earnings forecasts to recent StockTwits activity. Estimations are made at the analyst-firm-day level. The sample includes all U.S. firms covered by at least one analyst between 2009 and 2017. The dependent variable is a binary variable equal to one if the analyst issues a new forecast (or a revision) on a given firm on day t and zero otherwise. #Messages is the number of StockTwits messages posted about a firm from $t-30$ to $t-1$. The number of messages is set to zero when the firm is not covered/discussed on StockTwits. Trading Volume is the total volume of trading on from $t-30$ to $t-1$. In Column (3), we impose that no news (from the Capital IQ Key Developments dataset) is released about the firm during the day (otherwise the observation is removed from the sample). In Column (4), we impose that no news is released about the firm from $t-30$ to t (otherwise the observation is removed from the sample). t-statistics in parentheses are based on standard errors clustered by firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Binary Variable (New Forecast=1)			
OLS:	(1)	(2)	(3)	(4)
# Messages	0.02***	0.03***	0.06***	0.06***
	(2.97)	(4.29)	(8.82)	(2.70)
Trading Volume		-0.13***	-0.04***	0.09*
		(-9.74)	(-4.12)	(1.86)
Analyst × Firm FE	Yes	Yes	Yes	Yes
Analyst × Date FE	Yes	Yes	Yes	Yes
Sample without news in Key Dev. at t	No	No	Yes	No
Sample without news in Key Dev. over $_{t-30\to t}$	No	No	No	Yes
N	80,434,931	80,379,362	69,414,958	3,147,979

Table A3: Social Media Data and Analysts' Recommendations

This table presents OLS estimates of the sensitivity of analysts' recommendations to the number of "Bullish" and "Bearish" ratings issued by StockTwits users. Estimations are made at the analyst-firm-day level. The sample includes all U.S. firms covered by at least one analyst between 2009 and 2017. The dependent variable is the last available recommendation made by analyst i on firm j at t (measured by the item ireccd in I/B/E/S and multiplied by -1 so that greater values of ireccd indicate better recommendations). *Rating* is the difference between the fraction of "Bullish" users and that of "Bearish" users about j at $t-1$. *Rating* is naturally bounded between -1 (all users are "Bearish") and +1 (all users are "Bullish"). We require that there are at least 10 users with an active rating about j. A rating is active if it is the last available rating, and if it is not stale at $t-1$. A rating is stale after 365 days. In Column (2), we impose that no news (from the Capital IQ Key Developments dataset) is released about the firm during the day (otherwise the observation is removed from the sample), i.e., at t. In Column (3), we impose that no news is released about the firm from $t-30$ to t (otherwise the observation is removed from the sample). t-statistics in parentheses are based on standard errors clustered by firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Analyst Recommendation		
OLS:	(1)	(2)	(3)
Rating	0.11***	0.11***	0.14***
	(6.89)	(7.13)	(4.22)
Analyst × Firm FE	Yes	Yes	Yes
Analyst × Date FE	Yes	Yes	Yes
Sample without news in Key Dev. at t	No	Yes	No
Sample without news in Key Dev. over $_{t-30\to t}$	No	No	Yes
N	33,758,191	28,677,022	879,011

8 Actual Messages vs. Hypothetical Messages

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

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

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

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

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

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

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

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

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

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

Although our presentation focuses on the case where analyst i follows only firm j, the source of variation that our test relies upon is the same when analysts cover several firms, if coverage is stable. Since coverage is persistent on average, most changes in w'_i (i.e., the average of $\overline{w_j}$ across the covered firms j) will be captured by the analysts' fixed effects η_i, and the main source of variation will come from the *aggregate* variation in the number of messages (and from the staggered deployment of the platform). To mitigate the concern that changes in analyst coverage (i.e., change in w'_i over time) could explain our results, we verify and show that our estimates are not materially affected when we filter out the variations in coverage using a Bartik-like instrument (see Section VI D.2 and Table VII in the paper), or when we focus on the sub-sample of analysts covering always the same firms (see Table A11 in Section 14 of this Appendix).

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

9 Do Our Measures Correlate with News from Standard Sources?

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

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

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

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

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

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

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

This table presents OLS estimates of the sensitivity of different measures of social media data exposure to news arrival from standard sources. Estimations are made at the firm-day level. The sample includes all U.S. firms that have been discussed at least once on StockTwits between 2009 and 2017, and that are covered by at least one analyst. In columns (1) to (3), *#Watchlist* is the number of StockTwits users having the firm in their watchlist on day t. In columns (4) to (6), *#Hypothetical Messages* is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (7) to (9), *#Messages* is the number of actual messages posted about the firm from $t-30$ to $t-1$. $\#\text{News}_t$ is the number of distinct news about the firm reported in Capital IQ Key Developments on day $_t$. $\#News_{t\to T}$ is the number of distinct news about the firm reported in Capital IQ Key Developments between day t and day T. Capital IQ Key Developments is a dataset providing structured summaries of material news and events for more than 800,000 firms worldwide. It monitors more than 230 categories of news (i.e., a "key development" item) including for example companies SEC filings, executive changes, M&A announcements, earnings announcements, changes in corporate guidance, delayed filings, SEC inquiries, or credit rating changes. Each "key development item" includes announced date, headline, situation summary, type, company role, and company identifiers. t-statistics in parentheses are based on standard errors clustered by firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	*#Watchlist*			*#Hypothetical Messages*			*#Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\#\text{News}_t$	-4.66			-2.82			5.67***		
	(-0.59)			(-0.82)			(2.97)		
$\#\text{News}_{t-1}$		-3.98			-1.95			9.11***	
		(-0.51)			(-0.59)			(4.65)	
$\#\text{News}_{t-30\to t-1}$			-2.73			-1.68			10.06***
			(-0.41)			(-0.61)			(5.91)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	18,664,998	18,661,528	18,560,734	18,664,998	18,661,528	18,560,734	18,664,998	18,661,528	18,560,734

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

This table presents OLS estimates of the sensitivity of different measures of social media data exposure to news arrival from standard sources. Estimations are made at the firm-day level. The sample includes all U.S. firms that have been discussed at least once on StockTwits between 2009 and 2017, and that are covered by at least one analyst. In columns (1) to (3), *#Watchlist* is the number of StockTwits users having the firm in their watchlist on day t. In columns (4) to (6), *#Hypothetical Messages* is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (7) to (9), *#Messages* is the number of actual messages posted about the firm from $t-30$ to $t-1$. Market Response to $\#\text{News}_t$ is the Absolute (value of the) Cumulative Abnormal Return ($ACAR_{j,t}$) observed in response to news about firm j reported in Capital IQ Key Developments on day t. Market Response to $\#News_t$ is set to zero when no news is reported. The cumulative abnormal return at t is computed with a two-day window $[t+0, t+1]$, using CRSP value-weighted index as a benchmark. Market Response to $\#News_{t\to T}$ is sum of all $ACAR_{j,t}$ observed in response to each news event about j reported in Capital IQ Key Developments between day t and day T. This variable is set to zero when no news is reported between t and T. Capital IQ Key Developments is a dataset providing structured summaries of material news and events for more than 800,000 firms worldwide. It monitors more than 230 categories of news (i.e., a "key development" item) including for example companies SEC filings, executive changes, M&A announcements, earnings announcements, changes in corporate guidance, delayed filings, SEC inquiries, or credit rating changes. Each "key development item" includes announced date, headline, situation summary, type, company role, and company identifiers. t-statistics in parentheses are based on standard errors clustered by firm. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	*#Watchlist*			*#Hypothetical Messages*			*#Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mkt Resp. to $\#\text{News}_t$	1.35			-0.44			5.14***		
	(0.80)			(-0.44)			(6.38)		
Mkt Resp. to $\#\text{News}_{t-1}$		1.60			-0.32			6.56***	
		(0.95)			(-0.33)			(7.74)	
Mkt Resp. to $\#\text{News}_{t-30\to t-1}$			-0.30			-0.67			4.91***
			(-0.26)			(-0.98)			(10.31)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902

10 Long-run Evolution by Industry

This appendix investigates whether the long-run evolution documented in Table II (Section V.A), and Table III (Section V.B) concentrates in specific industries, using Fama-French industry classifications.

First, we focus on the level of R^2 by horizon and by industry. For each sub-sample, we estimate the trend in R^2 by regressing R^2 on a year counter variable, denoted "Year Trend", and report the estimated coefficient on this year counter with its t-statistic in Columns 1 to 5 of Table A6. For horizons shorter than one year and two years (greater than two years), the estimated trend by Fama-French industry is generally positive (negative). For example, 95% of the estimated coefficients on "Year Trend" are positive when $0 < h \leq 1$. By contrast, 79% are negative when $4 < h \leq 5$, in line with what we observe at the aggregate level.

Next, we focus on the slope of the term structure by industry. For each industry, we estimate the slope every year as in Table III (Section V.B) Column 1. Then, we estimate the trend in the slope of the term structure by regressing once again the slope on the year counter variable "Year Trend", and report the estimated coefficient and t-statistic in Table A6, Column 6. 76% of the estimated trends are negative indicating that the term structure has become steeper over time for a majority of industries.

Overall, the evidence from these tests suggests that the increase (decrease) in short-term (long-term) R^2 documented in Table II (Section V.A), and the simultaneous steepening of the term structure documented in Table III (Section V.B) are broad-based and robust.

Table A6: Long-Run Evolution by Industry

This table presents OLS estimates of time trend in analysts' forecasts' informativeness (columns 1 to 5) and in the slope of the term structure (column 6) by Fama-French industry. In columns (1) to (5), 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. In column (6), the dependent variable is the slope of the term-structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. The slope is calculated every year by regressing the average of R^2 by horizon on the horizon h. 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. For each industry-horizon sub-sample (columns 1 to 5) and each industry (column 6), we regress the dependent variable on a year counter (denoted"Year Trend") and a constant, and report the estimated regression coefficient on "Year Trend" with its t-statistic. "Year Trend" takes the value of zero for the period 1983-1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative change in R^2 (columns 1 to 5) and in slope (column 6) over the 1990-2017 period. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period (columns 1 to 5) and by year (column 6). In columns 1 to 5, we require that each OLS estimate for "Year Trend" is based on a sample of observations with at least 20 distinct analysts. In column 6, we require that each OLS estimate for "Year Trend" is based on a sample with at least 10 observations. Variable definitions are in Appendix II. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)										Slope by year	
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$3 < h \leq 4$		$4 < h \leq 5$		All years	
OLS	(1)		(2)		(3)		(4)		(5)		(6)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Panel A: Fama-French 5 Industries												
Consumer	14.7***	(11.75)	9.5***	(3.97)	4.5	(1.37)	-2.4	(-0.35)	-15.3**	(-2.03)	-9.9***	(-3.73)
Manufacturing	13.2***	(5.93)	6.0***	(2.45)	-2.8	(-0.77)	-13.1***	(-3.81)	-10.7**	(-1.98)	-6.9***	(-4.72)
Business Equipment	17.8***	(13.68)	22.7***	(10.79)	13.3***	(3.90)	2.9	(0.37)	6.4	(0.80)	-7.8**	(-2.36)
Healthcare	6.0***	(2.81)	3.2	(1.34)	-4.1	(-1.07)	-12.8*	(-1.94)	-27.7***	(-3.51)	-9.5***	(-2.88)
Other	7.9***	(4.54)	7.0***	(3.49)	5	(1.44)	-16.8**	(-2.22)	-32.2***	(-5.27)	-5.2**	(-1.97)
Panel B: Fama-French 12 Industries												
Consumer Nondurables	9.7***	(6.77)	1	(0.39)	-6.4*	(-1.78)	7.3	(0.92)	1.4	(0.14)	-3.6	(-1.04)
Consumer Durables	13.4***	(4.40)	5.7	(1.30)	-2.6	(-0.44)	-14	(-1.04)	-23.5**	(-2.14)	-14.0***	(-4.03)
Manufacturing	13.9***	(6.53)	8.6***	(3.32)	2.4	(0.58)	-5.7	(-0.82)	-5.3	(-0.55)	-6.4***	(-2.71)
Energy	16.3***	(4.33)	6.8	(1.24)	-8.5	(-1.07)	-23.2***	(-3.35)	-25.2***	(-2.70)	-8.3***	(-2.78)
Chemicals	10.6***	(4.34)	3.2	(0.82)	3.4	(0.44)	-4.2	(-0.32)	-1.6	(-0.07)	3.9	(1.22)
Business Equipment	19.3***	(14.84)	27.8***	(10.24)	24.1***	(5.81)	5.8	(0.67)	8.2	(0.75)	-10.3***	(-2.59)
Telecom	11.8***	(3.46)	9.2***	(2.51)	1.9	(0.31)	-1.4	(-0.15)	4.3	(0.28)	-2.6	(-1.16)
Utilities	17.2***	(4.79)	14.3***	(3.67)	-2.7	(-0.46)	-5.7	(-0.74)	-2.3	(-0.22)	-2.6	(-1.13)
Shops	18.3***	(12.80)	16.1***	(5.92)	17.7***	(3.78)	7.5	(0.82)	-5.3	(-0.35)	-9.1***	(-3.16)
Health	6.1***	(2.87)	3.4	(1.39)	-4	(-1.07)	-12.7*	(-1.93)	-27.8***	(-3.50)	-9.4***	(-2.90)
Finance	2	(0.66)	1.2	(0.26)	8.8	(1.27)	-19.6	(-1.41)	-49.9***	(-3.67)	-6.0*	(-1.69)
Other	10.3***	(5.76)	10.6***	(6.35)	6.0*	(1.63)	-8.9	(-1.03)	-19.6*	(-1.85)	-1.3	(-0.49)
Panel C: Fama-French 17 Industries												
Food	10.0***	(5.86)	4.4	(1.59)	-0.1	(-0.03)	-0.1	(-0.01)	-15.8	(-1.17)	-3.9	(-0.79)
Mines	-5	(-1.41)	-13.6**	(-2.38)	-12.8	(-1.45)	0.3	(0.03)	-19.9	(-0.76)	6.4*	(1.75)
Oil	16.5***	(4.30)	7.4	(1.30)	-7.7	(-0.88)	-23.3***	(-3.27)	-24.7***	(-2.54)	-8.7***	(-2.95)
Clothes	17.1***	(6.57)	10.3**	(2.24)	10.8	(0.77)	108.3***	(2.83)	-	-	8.1	(1.14)
Consumer Durables	12.6***	(4.59)	8.5*	(1.89)	1.1	(0.11)	22.3	(1.56)	-	-	-0.1	(-0.02)
Chemicals	8.5***	(2.65)	-2.6	(-0.58)	1.8	(0.23)	-9.9	(-0.80)	-9.9	(-0.45)	3.5	(0.91)
Consumer Nondurables	5.2***	(2.39)	-2.8	(-1.19)	-14.2***	(-3.19)	-26.3***	(-3.66)	-40.7***	(-4.12)	-9.1***	(-3.07)
Construction	8.3***	(3.02)	3.7	(0.76)	4.6	(0.52)	7.2	(0.53)	-18.5	(-1.11)	-5.7*	(-1.90)

Table A6: Long-Run Evolution by Industry (Cont'd)

This table presents OLS estimates of time trend in analysts' forecasts' informativeness (columns 1 to 5) and in the slope of the term structure (column 6) by Fama-French industry. In columns (1) to (5), 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. In column (6), the dependent variable is the slope of the term-structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. The slope is calculated every year by regressing the average of R^2 by horizon on the horizon h. 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. For each industry-horizon sub-sample (columns 1 to 5) and each industry (column 6), we regress the dependent variable on a year counter (denoted"Year Trend") and a constant, and report the estimated regression coefficient on "Year Trend" with its t-statistic. "Year Trend" takes the value of zero for the period 1983-1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative change in R^2 (columns 1 to 5) and in slope (column 6) over the 1990-2017 period. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period (columns 1 to 5) and by year (column 6). In columns 1 to 5, we require that each OLS estimate for "Year Trend" is based on a sample of observations with at least 20 distinct analysts. In column 6, we require that each OLS estimate for "Year Trend" is based on a sample with at least 10 observations. Variable definitions are in Appendix II. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)										Slope by year	
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$3 < h \leq 4$		$4 < h \leq 5$		All years	
OLS	(1)		(2)		(3)		(4)		(5)		(6)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Panel C: Fama-French 17 Industries (Cont'd)												
Steel	12.4**	(2.26)	-0.6	(-0.12)	-6.8	(-0.93)	-1.2	(-0.08)	-	-	1.4	(0.31)
Fabricated Products	18.0***	(3.02)	9	(1.15)	-14.1	(-0.73)	-5	(-0.33)	-	-	-	-
Machinery	17.3***	(10.79)	17.7***	(6.46)	8	(1.56)	-14.8	(-1.54)	-10.6	(-0.94)	-12.4***	(-2.93)
Automotive	13.6***	(4.09)	6.3	(1.31)	0.6	(0.09)	-1.1	(-0.09)	-6.6	(-0.56)	-9.3***	(-2.63)
Transportation	16.2***	(7.43)	13.9***	(5.35)	9.1***	(2.76)	8.3	(0.95)	7.9	(0.93)	-2.3	(-0.91)
Utilities	17.1***	(4.83)	13.8***	(3.52)	-3	(-0.50)	-5.7	(-0.74)	-2.9	(-0.27)	-2.5	(-1.08)
Retail Stores	19.1***	(13.13)	16.5***	(5.38)	15.9***	(3.28)	5	(0.52)	-8.6	(-0.52)	-9.8***	(-3.19)
Finance	2.8	(0.88)	2.2	(0.45)	8.4	(1.19)	-19.9	(-1.43)	-49.9***	(-3.67)	-6.2*	(-1.76)
Other	14.9***	(11.15)	19.2***	(11.92)	14.2***	(6.23)	-0.4	(-0.10)	-7	(-1.05)	-7.6***	(-5.29)
Panel D: Fama-French 48 Industries												
Agriculture	6.5	(0.92)	-3.9	(-0.36)	136.1*	(1.77)	-	-	-	-	-	-
Food Products	12.1***	(5.87)	5.8	(1.59)	5.9	(0.82)	22.7	(1.60)	-	-	-2.5	(-0.53)
Candy & Soda	6.4***	(2.85)	4.4	(0.91)	-13.1	(-0.77)	-66.0*	(-1.93)	-	-	-3.2	(-0.44)
Beer & Liquor	1.6	(0.32)	-6.9	(-0.81)	-14.5	(-1.21)	-55.5**	(-2.36)	-	-	-7.7	(-1.31)
Tobacco Products	10.8***	(5.53)	13.2***	(2.47)	13.5*	(1.82)	12.3	(0.76)	-	-	2.4	(0.47)
Recreation	15.8***	(3.03)	7.5	(1.11)	-8.9	(-0.69)	-	-	-	-	-15.4*	(-2.11)
Entertainment	14.4***	(3.28)	18.0***	(2.72)	12.1	(1.17)	-24.4*	(-1.85)	-	-	-6.2	(-1.46)
Printing and Publishing	9.7**	(2.23)	-0.3	(-0.04)	-23.6**	(-2.10)	-	-	-	-	-13.2**	(-2.25)
Consumer Goods	14.5***	(6.97)	11.7**	(2.33)	3.3	(0.24)	19.8	(0.72)	-	-	-1.5	(-0.14)
Apparel	14.3***	(4.93)	11.5**	(2.14)	19	(1.03)	-	-	-	-	3.3	(0.32)
Healthcare	16.7***	(4.81)	29.7***	(5.74)	30.9***	(3.63)	25.2	(1.42)	-	-	-0.4	(-0.09)
Medical Equipment	19.9***	(8.29)	34.1***	(8.57)	46.3***	(4.64)	26.8	(1.35)	-22.5	(-0.45)	3.8	(0.42)
Pharmaceutical Products	-0.4	(-0.17)	-10.3***	(-4.40)	-15.8***	(-4.33)	-16.8***	(-2.56)	-28.5***	(-3.52)	-8.7***	(-2.60)
Chemicals	8.4***	(2.90)	-2.3	(-0.53)	1	(0.12)	-8.8	(-0.81)	-10	(-0.46)	3.6	(0.96)
Rubber and Plastic Products	20.7***	(3.30)	35.8***	(3.26)	-2.8	(-0.11)	-	-	-	-	-7	(-0.53)
Textiles	8.5	(1.02)	-25.4**	(-1.93)	-21.4	(-1.31)	-	-	-	-	9.4	(0.73)
Construction Materials	12.0***	(4.02)	11.4***	(2.41)	2.9	(0.27)	-	-	-	-	-8.2**	(-2.00)
Construction	6.5*	(1.63)	-2.5	(-0.35)	17.5	(1.47)	15.3	(0.98)	-	-	0.3	(0.05)
Steel Works Etc	13.3***	(2.55)	1.5	(0.29)	-7.1	(-1.03)	-1.1	(-0.07)	-	-	0.9	(0.21)
Fabricated Products	8.4	(0.82)	28.0***	(2.98)	-	-	-	-	-	-	-	-

Table A6: Long-Run Evolution by Industry (Cont'd)

This table presents OLS estimates of time trend in analysts' forecasts' informativeness (columns 1 to 5) and in the slope of the term structure (column 6) by Fama-French industry. In columns (1) to (5), 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. In column (6), the dependent variable is the slope of the term-structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. The slope is calculated every year by regressing the average of R^2 by horizon on the horizon h. 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. For each industry-horizon sub-sample (columns 1 to 5) and each industry (column 6), we regress the dependent variable on a year counter (denoted"Year Trend") and a constant, and report the estimated regression coefficient on "Year Trend" with its t-statistic. "Year Trend" takes the value of zero for the period 1983-1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative change in R^2 (columns 1 to 5) and in slope (column 6) over the 1990-2017 period. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period (columns 1 to 5) and by year (column 6). In columns 1 to 5, we require that each OLS estimate for "Year Trend" is based on a sample of observations with at least 20 distinct analysts. In column 6, we require that each OLS estimate for "Year Trend" is based on a sample with at least 10 observations. Variable definitions are in Appendix II. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)										Slope by year	
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$3 < h \leq 4$		$4 < h \leq 5$		All years	
OLS	(1)		(2)		(3)		(4)		(5)		(6)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Panel D: Fama-French 48 Industries (Cont'd)												
Machinery	18.0***	(8.72)	16.4***	(4.56)	5.1	(0.65)	-23.3	(-1.45)	4.8	(0.32)	-3.6	(-0.63)
Electrical Equipment	16.8***	(4.75)	15.7***	(2.41)	11.4	(0.76)	-9.9	(-0.31)	-	-	-17.1***	(-2.71)
Automobiles and Trucks	12.9***	(3.72)	4.9	(0.95)	-1.8	(-0.28)	-6.7	(-0.49)	-14.6	(-1.24)	-9.8***	(-2.74)
Aircraft	9.8***	(2.98)	5.3	(1.03)	1.2	(0.15)	0.2	(0.02)	-1.6	(-0.12)	-1.3	(-0.39)
Shipbuilding, Railroad Eq.	8.5	(0.92)	-26.3*	(-1.81)	-	-	-	-	-	-	-	-
Defense	10.1	(1.36)	-23.1*	(-1.72)	3.6	(0.16)	-	-	-	-	-	-
Precious Metals	-8.8**	(-2.26)	-18.0***	(-2.79)	-13.9*	(-1.67)	-11.2	(-0.82)	-8.9	(-0.40)	6.6	(1.48)
Metal Mining	-13.4***	(-2.87)	-15.0*	(-1.92)	-25	(-1.09)	46.2**	(2.39)	-	-	-1.1	(-0.15)
Coal	28.3***	(3.02)	6.4	(0.56)	-6.1	(-0.20)	-	-	-	-	-	-
Petroleum and Natural Gas	16.0***	(4.20)	6.9	(1.21)	-7.8	(-0.90)	-23.2***	(-3.28)	-26.7***	(-2.87)	-8.4***	(-2.83)
Utilities	17.2***	(4.86)	14.1***	(3.69)	-2.7	(-0.45)	-5.7	(-0.74)	-2.3	(-0.22)	-2.5	(-1.10)
Communication	11.5***	(3.44)	8.5**	(2.38)	1.7	(0.27)	-0.4	(-0.04)	4.3	(0.28)	-2.6	(-1.16)
Personal Services	18.9***	(5.80)	20.3***	(4.27)	-26.7*	(-1.93)	-	-	-	-	-	-
Business Services	18.3***	(11.77)	31.5***	(10.04)	32.8***	(7.64)	21.9**	(2.29)	21.1*	(1.68)	4.3	(1.61)
Computers	19.0***	(7.99)	18.0***	(4.46)	3.2	(0.46)	-26.1	(-1.57)	-18	(-0.92)	-21.4***	(-5.86)
Electronic Equipment	16.4***	(7.08)	18.4***	(4.88)	5.4	(0.83)	9.4	(0.67)	13.6	(0.77)	-5.2	(-1.55)
Measuring and Control Eq.	18.2***	(6.57)	32.8***	(8.58)	51.0***	(5.12)	71.8***	(4.40)	-	-	26.1***	(5.42)
Business Supplies	17.0***	(5.15)	4.5	(0.84)	1.9	(0.22)	-21.6	(-1.05)	-	-	-12.7***	(-2.59)
Shipping Containers	4.4	(1.17)	-5.4	(-0.62)	0.7	(0.04)	-	-	-	-	22.2*	(1.79)
Transportation	19.7***	(7.21)	19.4***	(5.62)	8.2	(1.04)	13.1	(0.84)	3.1	(0.21)	1.9	(0.32)
Wholesale	11.6***	(4.87)	13.2***	(3.33)	28.1***	(2.95)	9.4	(0.21)	-	-	6.5	(1.02)
Retail	18.5***	(10.42)	15.3***	(4.83)	16.6***	(2.92)	7.7	(0.76)	-6.9	(-0.40)	-7.3***	(-2.86)
Restaurants, Hotels, Motels	22.2***	(7.29)	26.9***	(3.38)	21.8	(1.58)	2.4	(0.10)	-	-	-10.6*	(-1.91)
Banking	0.7	(0.20)	-2.6	(-0.39)	-4.9	(-0.36)	-	-	-	-	-7.1	(-1.30)
Insurance	0.3	(0.07)	-1.7	(-0.28)	9.8	(1.25)	-6.9	(-0.48)	-45.0**	(-2.25)	-2.2	(-0.62)
Real Estate	8.9	(1.27)	63.4***	(6.00)	-	-	-	-	-	-	-	-
Trading	6.7**	(2.16)	8.2	(1.38)	1.9	(0.17)	-50.3***	(-2.93)	-131.5***	(-6.57)	-14.0**	(-2.37)
Almost Nothing	13.0***	(6.94)	15.9***	(5.51)	13.4**	(2.11)	5.4	(0.32)	-16.1	(-1.22)	-1.9	(-0.61)

11 Long-run Evolution by Firm and Analyst Characteristics

This appendix investigates whether the long-run evolution documented in Table II (Section V.A), and Table III (Section V.B) concentrates in groups of firms with specific characteristics in terms of size, growth, exposure to high-technology industries, institutional ownership, share turnover, and option trading, or in groups of analysts with specific experience.

First, we focus on the level of R^2 by horizon and by group. For each sub-sample, we estimate the trend in R^2 by regressing R^2 on a year counter variable, denoted "Year Trend", and report the estimated coefficient on this year counter with the associated t-statistic in Table A6, Columns 1 to 5. For horizons shorter than one year (greater than three years), the estimated trend is positive (negative) regardless of how we form groups.

Next, we focus on the slope of the term structure by group. For each group, we estimate the slope every year as in Table III (Section V.B) Column 1. Then, we estimate the trend in the slope of the term structure by regressing the slope on the same year counter variable "Year Trend", and again report the estimated coefficient and t-statistic in Table A6, Column 6. All estimated trends are negative indicating that the term structure has become steeper over time for all groups of firms and analysts.

Overall, the evidence from these tests indicates that the aggregate shift in the term structure of analysts' forecasts informativeness over time is not driven by a specific group of firms and analysts. This trend is broad-based and robust to various data sampling.

Table A7: Long-Run Evolution by Firm and Analyst Characteristics

This table presents OLS estimates of time trend in analysts' forecasts' informativeness (columns 1 to 5) and in the slope of the term structure (column 6) by group of firms and analysts characteristics. In columns (1) to (5), 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. In column (6), the dependent variable is the slope of the term-structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. The slope is calculated every year by regressing the average of R^2 by horizon on the horizon h. 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. For each group or group-horizon sub-sample, we regress the dependent variable on a year counter (denoted"Year Trend") and a constant, and report the estimated regression coefficient on "Year Trend" with its t-statistic. In all panels except panel H,"Year Trend" takes the value of zero for the period 1983-1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative change in R^2 (columns 1 to 5) and in slope (column 6) over the 1990-2017 period. In panel H, "Yes" indicates that all stocks covered by the analyst have traded options, and "No" that at least one covered stock does not have traded options. Because information about option listing is available from 1996 onwards, "Year Trend" takes the value of zero for the period 1996-2001 and increments by one every subsequent year, divided by 16. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period (columns 1 to 5) and by year (column 6). In columns 1 to 5, we require that each OLS estimate for "Year Trend" is based on a sample of observations with at least 20 distinct analysts. In column 6, we require that each OLS estimate for "Year Trend" is based on a sample with at least 20 observations. Variable definitions are in Appendix II. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)										Slope by year	
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$3 < h \leq 4$		$4 < h \leq 5$		All years	
OLS	(1)		(2)		(3)		(4)		(5)		(6)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Panel A: Low vs. High Total Assets (Inflation adjusted)												
Below Median	12.5***	(10.12)	9.9***	(6.31)	5.7**	(2.09)	-1.3	(-0.49)	-24.1***	(-5.27)	-10.5***	(-6.37)
Above Median	12.7***	(6.73)	9.0***	(4.78)	-0.3	(-0.15)	-14.2***	(-4.39)	-15.2***	(-3.05)	-10.2***	(-5.30)
Panel B: Low vs. High Market Cap. (Inflation adjusted)												
Below Median	12.1***	(7.48)	7.2***	(5.07)	0.9	(0.35)	-5.2	(-1.56)	-20.6***	(-3.95)	-10.4***	(-5.81)
Above Median	10.3***	(6.67)	7.6***	(4.28)	-0.3	(-0.13)	-17.0***	(-6.22)	-22.7***	(-6.09)	-10.8***	(-6.12)
Panel C: Low vs. High Sales Growth												
Below Median	12.9***	(6.72)	8.3***	(4.36)	1.5	(0.64)	-10.4***	(-3.78)	-15.2***	(-3.33)	-9.8***	(-4.82)
Above Median	12.7***	(9.76)	11.2***	(5.60)	3.7*	(1.63)	-8.7***	(-2.39)	-21.9***	(-4.42)	-9.8***	(-5.52)
Panel D: Low vs. High Book-to-Market												
Below Median	10.6***	(9.41)	12.2***	(4.83)	8.1***	(2.55)	-11.3***	(-2.98)	-24.3***	(-4.54)	-11.0***	(-6.91)
Above Median	12.7***	(6.53)	7.0***	(3.45)	-3.6	(-1.03)	-10.2***	(-2.70)	-13.8***	(-2.70)	-10.3***	(-4.99)
Panel E: Tech. vs. Non-tech. Firms												
Non-tech. Firms	15.4***	(13.64)	17.8***	(10.02)	5.2**	(1.99)	-9.2*	(-1.86)	-15.6**	(-2.38)	-8.8***	(-5.26)
Tech. Firms	11.9***	(7.33)	7.6***	(4.49)	2.3	(0.93)	-8.4***	(-2.89)	-14.7***	(-3.86)	-12.4***	(-4.40)

Table A7: Long-Run Evolution by Firm and Analyst Characteristics (Cont'd)

This table presents OLS estimates of time trend in analysts' forecasts' informativeness (columns 1 to 5) and in the slope of the term structure (column 6) by group of firms and analysts characteristics. In columns (1) to (5), 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. In column (6), the dependent variable is the slope of the term-structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. The slope is calculated every year by regressing the average of R^2 by horizon on the horizon h. 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. For each group or group-horizon sub-sample, we regress the dependent variable on a year counter (denoted"Year Trend") and a constant, and report the estimated regression coefficient on "Year Trend" with its t-statistic. In all panels except panel H,"Year Trend" takes the value of zero for the period 1983-1989 and increments by one every subsequent year, divided by 28 so that the regression coefficient can be interpreted as the cumulative change in R^2 (columns 1 to 5) and in slope (column 6) over the 1990-2017 period. In panel H, "Yes" indicates that all stocks covered by the analyst have traded options, and "No" that at least one covered stock does not have traded options. Because information about option listing is available from 1996 onwards, "Year Trend" takes the value of zero for the period 1996-2001 and increments by one every subsequent year, divided by 16. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period (columns 1 to 5) and by year (column 6). In columns 1 to 5, we require that each OLS estimate for "Year Trend" is based on a sample of observations with at least 20 distinct analysts. In column 6, we require that each OLS estimate for "Year Trend" is based on a sample with at least 20 observations. Variable definitions are in Appendix II. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)										Slope by year	
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$3 < h \leq 4$		$4 < h \leq 5$		All years	
OLS	(1)		(2)		(3)		(4)		(5)		(6)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Panel F: Low vs. High Institutional Ownership												
Below Median	9.0***	(4.44)	0.3	(0.14)	-10.5***	(-3.94)	-21.0***	(-7.80)	-30.4***	(-6.88)	-11.6***	(-6.54)
Above Median	7.6***	(7.42)	14.6***	(8.45)	13.7***	(4.17)	-4.2	(-0.86)	-10.9*	(-1.70)	-5.3	(-1.35)
Panel G: Low vs. High Share Turnover												
Below Median	12.6***	(7.90)	6.9***	(4.32)	-1.6	(-1.01)	-16.3***	(-6.22)	-24.6***	(-5.78)	-10.7***	(-6.67)
Above Median	14.6***	(10.26)	20.0***	(12.58)	13.5***	(4.71)	4.1	(1.16)	-5.5	(-1.20)	-7.4***	(-3.84)
Panel H: Option vs. No Option trading												
Yes	5.1***	(5.61)	9.1***	(6.80)	5.7***	(2.80)	-4.6*	(-1.65)	-7.7**	(-2.24)	-5.5***	(-3.62)
No	13.4***	(4.15)	25.7***	(5.21)	20.9*	(1.70)	0.5	(0.03)	-32.9*	(-2.03)	-	-
Panel I: Low vs. High Analyst Experience (# years in I/B/E/S since inception)												
Below Median	14.0***	(10.70)	8.8***	(7.05)	-1.5	(-0.82)	-14.6***	(-6.03)	-23.7***	(-5.47)	-12.3***	(-6.81)
Above Median	8.3***	(6.84)	11.5***	(6.17)	8.2***	(3.82)	-6.9*	(-1.79)	-16.6***	(-3.14)	-6.9***	(-4.03)
Panel J: Low vs. High Analyst Experience (# covered firms in I/B/E/S since inception)												
Below Median	12.0***	(8.31)	8.9***	(5.16)	1.9	(0.91)	-13.2***	(-4.63)	-20.2***	(-4.75)	-10.6***	(-6.69)
Above Median	13.9***	(9.95)	12.2***	(9.13)	4.5**	(2.34)	-6.2**	(-2.03)	-18.2***	(-4.20)	-11.0***	(-4.99)

12 Robustness Table II

This Appendix discusses the robustness of the results reported in Table II (Section V.A). All robustness tests are reported in Table A8.

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

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

Finally, we check that our results are not specific to using the period 1983-1989 as our baseline, nor driven by I/B/E/S imperfect coverage at the beginning of the sample (Panel F). We also show that neither the number of forecasts used to estimate R^2 (Panel G), nor the assumptions we make about the updating speed of those forecasts (Panel H), materially affects inferences. Panel G reports specifications that include fixed effects for the number of observations used to estimate R^2 in eq.(14). Panel H reports results after we re-compute R^2 assuming analysts constantly update their forecasts. Specifically, we estimate an updated forecast every day, unless the analyst discloses one. We do so by linear interpolation between two consecutive disclosures for each analyst, firm, and fiscal period. This alternative approach for computing R^2 relaxes the implicit assumption that analysts update their forecasts only when we observe a new forecast.

Table A8: Robustness: Forecast Informativeness by Horizon

This table presents OLS estimates of time trend in analysts' forecasts' informativeness by sub-samples including observations in different annual forecasting horizon ranges. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. h is the forecasting horizon, measured as the number of days between the forecasting date and the date of actual earnings release, divided by 365. Year Trend is a variable that takes the value of zero for the period 1983-1989 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 1990-2017 period. The constant is omitted when absorbed by the fixed effects. Variable definitions are in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	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)
Panel A: Controlling for 2-digit SIC Industry Fixed-Effects					
Year Trend	13.4***	11.6***	1.9	-8.2***	-15.7***
	(9.84)	(7.92)	(1.11)	(-3.18)	(-3.61)
SIC2 FE	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No
N	33,386,528	25,044,127	5,359,098	1,349,651	703,653
Panel B: Controlling for the characteristics of the portfolio of covered firms					
Year Trend	8.8***	7.5***	0.8	-6.0*	-9.2*
	(6.71)	(5.85)	(0.39)	(-1.71)	(-1.69)
SIC2 FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
N	31,158,684	23,203,643	4,992,082	1,285,671	670,019
Panel C: Focusing on SP500 firms					
Year Trend	13.2***	12.4***	6.4***	-4.2	-9.8**
	(7.31)	(6.13)	(2.76)	(-1.43)	(-2.05)
Constant (83-89)	79.0***	63.3***	55.7***	53.7***	50.4***
	(65.64)	(55.91)	(43.35)	(34.29)	(21.98)
N	18,423,237	14,206,102	3,138,963	769,951	406,058
Panel D: Analysts with short and long-term forecasts					
Year Trend	8.0***	6.8***	1.7	-12.4***	-21.8***
	(5.11)	(4.38)	(0.88)	(-5.11)	(-5.52)
Constant (83-89)	77.7***	57.7***	47.2***	45.3***	44.5***
	(77.09)	(56.61)	(30.17)	(27.17)	(19.38)
N	8,600,935	7,389,585	3,663,585	1,349,749	703,712

Table A8: Robustness: Forecast Informativeness by Horizon (Cont'd)

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)
Panel E: Firms with short and long-term forecasts					
Year Trend	8.7***	4.1***	0	-12.3***	-21.9***
	(5.49)	(2.49)	(-0.01)	(-5.04)	(-5.49)
Constant (83-89)	77.7***	60.4***	50.3***	45.5***	44.7***
	(77.48)	(65.83)	(38.33)	(27.00)	(19.25)
N	29,023,675	22,491,017	5,159,145	1,338,504	698,958
Panel F: Excluding 80's					
Year Trend	7.6***	8.5***	3.5*	-11.8***	-18.3***
	(6.19)	(5.54)	(1.72)	(-4.66)	(-4.78)
Constant (90-92)	77.4***	55.6***	47.1***	44.5***	41.4***
	(113.19)	(62.63)	(31.24)	(26.28)	(19.77)
N	29,047,461	22,334,402	5,169,002	1,308,876	683,413
Panel G: Controlling for the number of observations used to compute R^2					
Year Trend	13.1***	10.7***	6.6***	-9.0***	-20.0***
	(9.35)	(7.61)	(3.39)	(-3.34)	(-5.29)
#Firms FE	Yes	Yes	Yes	Yes	Yes
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712
Panel H: Using R^2 based on interpolated forecasts					
Year Trend	10.8***	7.4***	-1.5	-11.7***	-14.5***
	(7.51)	(5.57)	(-1.36)	(-5.26)	(-4.08)
Constant (83-89)	77.3***	60.4***	56.3***	54.4***	52.1***
	(95.76)	(101.00)	(65.69)	(35.84)	(23.58)
N	33,413,667	25,060,925	5,361,069	1,349,749	703,712

13 Robustness Table III

This Appendix discusses the robustness of the results reported in Table III (Section V.B). All robustness tests are reported in Table A9.

In Panel A, we report specifications controlling for the average characteristics of the covered firms, namely: size (log of total assets), (log of) age, cash flow to assets, debt to assets, cash to assets, Tobin's Q, sales growth, share turnover, institutional ownership, and the fraction of tech-firms covered. Specifically, we average those average characteristics by year in Column (1), by (two-digit SIC) industry and year in Columns (2) and (3), and by analyst and year in Columns (4) and (5), and control for those in the regression. Panel B shows similar results using similar specifications, but after we re-compute R^2 focusing only on forecasts about S&P500 firms, whose underlying characteristics have remained stable over time (Bai et al. (2016)).

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

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

Table A9: The Slope of the Term Structure

This table presents OLS estimates of time trend in the slope of the term structure of forecasts' informativeness. The dependent variable is the slope of the term structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. A negative slope indicates that forecasts' informativeness (R^2) decreases with horizon. In column (1), the slope is calculated every year by regressing the average of R^2 by horizon on the horizon h (i.e., the number of days between the forecasting date and the date of actual earnings release, divided by 365). In columns (2) and (3), the slope is calculated every year by 2-digit SIC industry by regressing the average of R^2 by horizon and industry on h. In columns (4) and (5), the slope is calculated every year by analyst by regressing the average of R^2 by horizon and analyst on h. Year Trend is a variable that takes the value of zero for the period 1983-1989 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 1990-2017 period. The constant is omitted when absorbed by the fixed effects. Variable definitions are in the Appendix II. t-statistics in parentheses are based on standard errors clustered by year. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS:	(1)	(2)	(3)	(4)	(5)
Panel A: Controlling for covered firms characteristics					
Year Trend	-19.9***	-4.1**	-7.1***	-3.4***	-5.3**
	(-3.72)	(-2.03)	(-3.64)	(-3.91)	(-2.21)
Constant (83-89)	-32.7	-19.0***			
	(-0.95)	(-4.76)			
Controls	Yes	Yes	Yes	Yes	Yes
SIC2 FE	-	No	Yes	Yes	Yes
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	3,899	3,440
Panel B: Focusing on SP500 firms					
Year Trend	-8.3***	-1.9	-2.8*	-5.6***	-3.8**
	(-3.67)	(-1.52)	(-1.91)	(-8.45)	(-2.14)
Constant (83-89)	-6.9***	-11.2***		-9.4***	
	(-4.82)	(-15.29)		(-21.28)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	803	772	4,533	4,307
Panel C: Analysts with short and long-term forecasts					
Year Trend	-10.8***	-5.0***	-3.4***	-5.4***	-3.0**
	(-6.31)	(-4.52)	(-2.87)	(-8.11)	(-2.12)
Constant (83-89)	-6.6***	-11.3***		-11.6***	
	(-6.05)	(-21.09)		(-23.71)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	7,657	7,290

Table A9: The Slope of the Term Structure (Cont'd)

Dep. variable:	Slope by year	Slope by SIC2-year		Slope by analyst-year	
OLS:	(1)	(2)	(3)	(4)	(5)
Panel D: Firms with short and long-term forecasts					
Year Trend	-10.0***	-4.2***	-2.9**	-4.9***	-2.8*
	(-5.70)	(-3.64)	(-2.37)	(-7.01)	(-1.86)
Constant (83-89)	-7.2***	-11.8***		-12.0***	
	(-6.38)	(-17.44)		(-22.86)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,050	1,019	7,619	7,252
Panel E: Excluding 80's					
Year Trend	-7.6***	-3.8***	-2.4**	-4.1***	-2.6*
	(-7.20)	(-4.11)	(-2.40)	(-5.20)	(-1.96)
Constant (90-92)	-8.5***	-12.0***		-12.7***	
	(-12.60)	(-23.92)		(-22.23)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	26	959	957	7,430	7,054
Panel F: Using R^2 based on interpolated forecasts					
Year Trend	-9.0***	-4.4***	-3.7***	-6.1***	-4.7***
	(-6.10)	(-4.77)	(-3.73)	(-8.57)	(-3.17)
Constant (83-89)	-4.7***	-9.0***		-8.4***	
	(-5.25)	(-19.06)		(-19.09)	
SIC2 FE	-	No	Yes	-	-
Analyst FE	-	-	-	No	Yes
N	33	1,083	1,080	7,657	7,290

14 Robustness Table VIII

This Appendix discusses the robustness of the results reported in Table VIII (Section VI.D). Table A10 shows that our results are robust to controlling for trading volume and thus for the effects of news (public and private) that are material enough for generating trading. Table A11 shows that our results are also robust to focusing on analysts with stable coverage, and thus that changes in coverage cannot be the main explanation for our findings. Finally, we verify that focusing on analysts (Table A12) and firms (Table A13) for which both short and long-term forecasts are available does not affect inferences. Table A12 repeats the analysis focusing on analysts who have issued at least one forecast with horizon greater than 3 years. Table A13 does the same, but after we re-calculate R^2 using only forecasts about firms for which at least one forecast with horizon greater than 3 years is available.

Table A10: Robustness: Controlling for Trading Volume

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(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 is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h-1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other control variables include firms' 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, Tobin's Q, and 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 $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. h^* is omitted when absorbed by the fixed effects. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-1.09***	-0.86***	-0.97***	-1.01***	-1.06***	-1.08***
	(-3.23)	(-3.17)	(-3.79)	(-3.88)	(-4.84)	(-5.15)
Data Exposure	0.16	-0.17	-0.26	0.38*	-0.14	-0.17
	(0.66)	(-0.68)	(-1.01)	(1.62)	(-0.62)	(-0.67)
$h^*\times$ Trading Volume	1.13***	0.62***	0.55***	1.18***	0.71***	0.64***
	(6.56)	(3.28)	(2.65)	(6.82)	(3.75)	(3.14)
Trading Volume	-0.4	-0.12	-0.82***	-0.43	-0.12	-0.83***
	(-1.29)	(-0.49)	(-2.64)	(-1.39)	(-0.48)	(-2.69)
h^*	-17.62***			-17.59***		
	(-31.69)			(-30.94)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,276	30,105,551	27,845,302	30,959,276	30,959,276	27,845,302

Table A11: Robustness: Analysts With Stable Coverage

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(19)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with stable coverage only. Coverage is stable if the level of similarity between the portfolio of firms covered in the current year and that of the previous year is greater than 90%. Similarity is defined as the number of common firms between the portfolio covered in the current year and the one covered the year before, scaled by the square root of the product of the number of firms in each portfolio. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other control variables include firms' 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, Tobin's Q, and 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 $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. h^* is omitted when absorbed by the fixed effects. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-0.70**	-0.61**	-0.74***	-0.62***	-0.87***	-0.94***
	(-2.20)	(-2.30)	(-2.77)	(-2.58)	(-3.98)	(-4.20)
Data Exposure	0.34	0.02	-0.08	0.50*	-0.03	-0.04
	(1.42)	(0.08)	(-0.29)	(1.84)	(-0.10)	(-0.15)
$h^*\times$ Trading Volume	1.05***	0.76***	0.63***	1.08***	0.84***	0.72***
	(5.80)	(3.47)	(2.58)	(5.87)	(3.69)	(2.89)
Trading Volume	-0.21	0.06	-0.47	-0.24	0.06	-0.47
	(-0.58)	(0.19)	(-1.28)	(-0.63)	(0.22)	(-1.27)
h^*	-17.33***			-17.32***		
	(-32.31)			(-31.82)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	14,552,288	13,773,488	12,678,809	14,552,288	13,773,488	12,678,809

Table A12: Robustness: Analysts With Non-Missing Long-Term Forecasts

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(19)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with non-missing long-term forecasts. An analyst has non-missing long-term forecasts if there is at least one non-missing $R^2_{i,t,h}$ for $h \geq 3$ over the sample period (2005-2017). The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other control variables include firms' 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, Tobin's Q, and 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 $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. h^* is omitted when absorbed by the fixed effects. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-1.60***	-1.15***	-1.26***	-1.38***	-1.32***	-1.29***
	(-4.44)	(-3.25)	(-4.21)	(-5.13)	(-5.74)	(-7.36)
Data Exposure	-0.06	-0.3	-0.44	0.24	-0.26	-0.27
	(-0.29)	(-1.03)	(-1.47)	(1.13)	(-1.13)	(-1.11)
$h^*\times$ Trading Volume	0.97***	0.64***	0.51**	1.05***	0.74***	0.61***
	(5.36)	(3.05)	(2.25)	(5.58)	(3.51)	(2.75)
Trading Volume	-0.46	-0.2	-0.57	-0.49*	-0.19	-0.58
	(-1.48)	(-0.63)	(-1.16)	(-1.62)	(-0.60)	(-1.19)
h^*	-16.24***			-16.19***		
	(-35.68)			(-34.38)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	13,782,999	13,019,477	12,149,315	13,782,999	13,019,477	12,149,315

Table A13: Robustness: Firms With Non-Missing Long-Term Forecasts

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(19)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts covering firms with non-missing long-term forecasts. A firm has non-missing long-term forecasts if it has at least one non-missing forecast for $h \geq 3$ over the sample period (2005-2017). The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other control variables include firms' 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, Tobin's Q, and 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 $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with horizon fixed effects. h^* is omitted when absorbed by the fixed effects. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecasts informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-0.75**	-0.43	-0.54*	-0.85***	-0.85***	-0.85***
	(-2.02)	(-1.38)	(-1.78)	(-2.88)	(-3.12)	(-3.42)
Data Exposure	0.55***	0.25	0.13	0.72***	0.22	0.16
	(2.47)	(0.94)	(0.49)	(3.50)	(0.93)	(0.69)
$h^*\times$ Trading Volume	1.00***	0.61***	0.56***	1.06***	0.69***	0.65***
	(5.97)	(3.11)	(2.51)	(6.17)	(3.48)	(2.87)
Trading Volume	-0.32	-0.03	-0.75**	-0.36	-0.04	-0.77**
	(-1.02)	(-0.10)	(-2.17)	(-1.15)	(-0.13)	(-2.24)
h^*	-17.31***			-17.27***		
	(-33.37)			(-32.58)		
Analyst Portfolio FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst Portfolio FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	29,690,225	28,896,486	26,695,116	29,690,225	28,896,486	26,695,116

15 Additional Predictions: Effects of multi-tasking costs (c) and earnings' auto-correlation (ρ)

In this section, we study how multitasking costs (c) and earnings auto-correlation (ρ) affect the sign (and magnitude) of the change in the informativeness of long-term forecasts in response to a change in the marginal cost of producing short term information (a). Corollary 1 implies that as ρ gets larger or c gets smaller, the effect of a decrease in a on the informativeness of the analyst's long-term forecast should switch from being negative to being positive. As explained in the text, this implies that the negative effect of a decrease in a on the informativeness of the long-term forecast should become weaker as ρ increases or c decreases, and could even potentially become positive when ρ is high enough and c small enough.

Table A14 illustrates this point by showing the *percentage change* in the informativeness of the long-term forecast (computed using eq.(13) for specific parameter values of the model and Proposition 1) when a decreases from 6 to 5. In this example we vary ρ from 0 to 0.7 by varying β from 0 to 1 ($\sigma_{st}^2 = \sigma_e^2 = 1$) and we vary c from 2 to 0 (c: Columns; β: lines). Other parameters are $\gamma = 0.5$ and $\psi_{st} = \psi_{lt} = 1$.

$c/\beta(\rho)$	2	1.75	1.5	1.25	1	0.75	0.5	0.25	0
0 (0)	-39.7	-27.7	-19.7	-14.1	-9.9	-6.6	-4	-1.8	0
0.25 (0.24)	-14.8	-10.3	-6.8	-4.1	-1.8	0.0	1.6	2.9	4
0.5 (0.44)	5.4	6.2	7	7.7	8.3	8.9	9.4	9.9	10.3

Table A14: Percentage change in the informativeness of the analyst's long-term forecast following a one unit decrease in a (from 6 to 5) when β varies from 0 to 1 (lines) and c varies from 2 to 0 (columns). Other parameter values are $\sigma_t^2 = \sigma_e^2 = 1$; $\gamma = 0.5$ and $\Psi_{st} = \Psi_{lt} = 1$. Values of ρ for each value of β are shown in parenthesis in the first column.

Clearly, as ρ increases and c decreases, the effect of a decrease in a on the informativeness of the long-term forecast becomes less negative and becomes positive at some point (compare the upper left part of the table with the lower right part).

This implies that the negative effect of a decrease in a on the slope of the term structure of analysts' forecast informativeness should become weaker as well when ρ increases or c

decreases. Table A15 illustrates this point for the same parameter values as in the previous table.

c/β (ρ)	2	1.75	1.5	1.25	1	0.75	0.5	0.25	0
0 (0)	-0.016	-0.015	-0.015	-0.015	-0.014	-0.014	-0.014	-0.014	-0.013
0.25 (0.24)	-0.015	-0.015	-0.014	-0.014	-0.014	-0.013	-0.013	-0.013	-0.013
0.5 (0.44)	-0.014	-0.013	-0.013	-0.013	-0.012	-0.012	-0.012	-0.012	-0.011

Table A15: Change in the slope of the term structure of analysts forecast informativeness (Δ in Corollary 1) following a unit decrease in a (from 6 to 5) when β varies from 0 to 1 and c varies from 2 to 0. Other parameter values are $\sigma_t^2 = \sigma_e^2 = 1$; $\gamma = 0.5$ and $\Psi_{st} = \Psi_{lt} = 1$. Values of ρ for each value of β are shown in parenthesis in the first column.

The pattern in Table A15 is consistent with the findings in Tables IX and X in the paper. That is, a decrease in the multitasking cost (Table IX) or an increase in the autocorrelation of earnings (Table X) weakens the negative effect of a reduction in a (proxied by analysts' exposure to StockTwits).

For completeness, Table A16 shows the results of estimating the effect of greater data exposure on R_h^2 with $2 < h \leq 3$ and $h > 3$ using eq.(17) (with and without controls), separately by tercile of multi-tasking cost (c) and auto-correlation (ρ). For brevity, we only report the estimated regression coefficient on "Data Exposure" sorted by tercile of c (as proxied by the number of covered stocks) and ρ. Overall, the negative effect of greater data exposure on R^2 increases with c holding ρ constant, and decreases with ρ holding c constant, as in Table A14 (note: in Table A16, variations in c are in lines while variations in ρ are in columns) . Also in line with theory, the sign of the effect changes and becomes positive when c is sufficiently small and ρ sufficiently large.

Table A16: Data Exposure and Forecast Informativeness by Cost of Multi-tasking (c) and Auto-correlation (ρ)

This table presents OLS estimates of eq.(17) by tercile of multi-tasking cost (c) and tercile of auto-correlation (ρ) separately for $2 < h \leq 3$ and $h \leq 3$. 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. We only report the regression coefficient on Data Exposure. We use *Coverage* as a proxy for c and Auto as proxy for ρ. *Coverage* is the number of firms that the analyst covers. Auto is the average earnings' autocorrelation in analysts' portfolios (measured using the R^2 of a regression of quarterly earnings on their lagged value). Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. Variable:	Forecasts informativeness (R2)					
Sample:	$2 < h \leq 3$			$h > 3$		

Panel A: Proxy for Data Exposure = #Watchlist (Specification without controls)

	Auto-correlation			Auto-correlation		
Multi-Tasking Cost	Low=1	2	High=3	Low=1	2	High=3
Low=1	0.27	2.13***	2.38**	3.53*	3.56***	1.96
2	-1.97***	0.12	-0.01	-1.16	-1.35**	-0.06
High=3	-2.23***	-1.35***	0.68	-2.68***	-1.91***	-0.85

Panel B: Proxy for Data Exposure = #Watchlist (Specification with controls)

	Auto-correlation			Auto-correlation		
Multi-Tasking Cost	Low=1	2	High=3	Low=1	2	High=3
Low=1	0.51	1.61***	1.95*	4.18**	4.49***	2.31*
2	-2.01***	-0.18	-0.3	-0.89	-0.93	0.13
High=3	-2.33***	-1.62***	-0.05	-2.82***	-1.90***	-1.02**

Panel C: Proxy for Data Exposure = #Hypothetical Messages (Specification without controls)

	Auto-correlation			Auto-correlation		
Multi-Tasking Cost	Low=1	2	High=3	Low=1	2	High=3
Low=1	1.23	2.49***	2.15	2.55	3.34***	2.51*
2	-1.52*	0.15	-0.04	-1.93**	-1.59*	-0.24
High=3	-2.84***	-1.83***	0.35	-3.47***	-2.73***	-1.51***

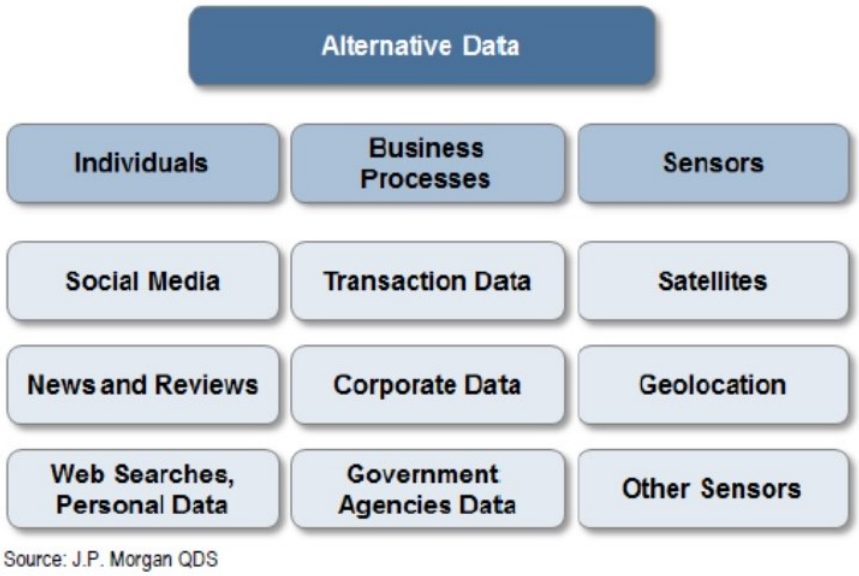
Panel D: Proxy for Data Exposure = #Hypothetical Messages (Specification with controls)

	Auto-correlation			Auto-correlation		
Multi-Tasking Cost	Low=1	2	High=3	Low=1	2	High=3
Low=1	1.84**	1.99***	1.7	2.96	4.14***	2.95***
2	-1.41*	-0.14	-0.31	-1.59*	-1.07	0.01
High=3	-2.75***	-1.96***	-0.22	-3.62***	-2.59***	-1.46**

16 Alternative Data: Definition and Classification

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

Figure 1: Classification of big/alternative data sources



Source: J.P. Morgan QDS

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

17 Example of Analysts Using Social Media Data

J.P.Morgan

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

AWS Partners Meeting Key Takeaways

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

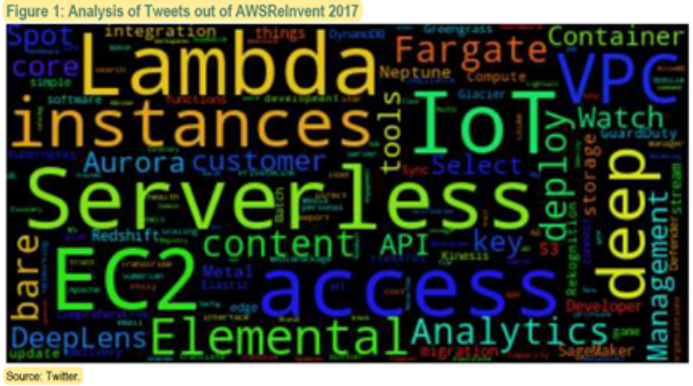
Internet
Doug Anmuth AC
(1-212) 622-6571
douglas.anmuth@jpmorgan.com
Bloomberg JPMA ANMUTH <GO>
Ashwin Kesireddy
(1-415) 315-6756
ashwin.x.kesireddy@jpmorgan.com
Cory A Carpenter
(1-212) 270-8125
cory.carpenter@jpmorgan.com
Software
Mark R Murphy AC

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

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

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

Figure 1: Analysis of Tweets out of AWSReInvent 2017



Source: Twitter.

18 Measuring Alternative Data Usage by Industry

This appendix explains how to construct the text-based measures of alternative data usage by industry used in Section V.C. The procedure consists of four steps:

- Step 1: Identify all alternative data vendors from the directory of alternative data vendors created by J.P.Morgan (2019 Alternative Data Handbook). The directory contains the name of 506 unique alternative data vendors (Pages 205-246). Separate names and acronyms, remove duplicates, and form keywords by adding the prefix "Source: " to all company names and acronyms. The list of keywords we obtain at the end of this first step is reported in Section 19 of this on-line appendix.

- Step 2: Search for all company reports by US analysts in Thomson-Reuters Refinitiv citing at least one alternative data vendor using each keyword with Eikon keyword search engine, and then download the list of match reports. The list includes the date of each report, the name of the analyst, and the ticker of the company. Drop duplicates and merge with CRSP by ticker and date to identify the Fama-French industry. Our final list contains 21,067 reports about 2,635 firms by 2,790 analysts over the 1983-2017 period citing one (or more) alternative data vendor as data source.

- Step 3: Count the number of reports (analysts) (firms) by industry-year and normalize by the total number of reports (analysts) (firms) in I/B/E/S in that industry-year to approximate the percentage of reports (analysts) (firms) using alternative data or covered by alternative data by industry-year.

- Step 4: Calculate the median percentage of reports (analysts) (firms) over the second half of our sample (2000-2017) by industry to identify (i) the frequency of alternative data references in company reports in that industry (Alternative Data Usage 1), (ii) the frequency of alternative data usage by analysts (Alternative Data Usage 2), and (iii) the abundance of alternative data coverage of firms from that industry (Alternative Data Usage 3).

Table A17 below shows descriptive statistics for all three measures using Fama-French 17 industry classification.

Table A17: Alternative Data Usage: Descriptive Statistics

	N	Mean	STDV	Min	P25	P50	P75	Max
Alternative Data Usage 1	17	0.2%	0.1%	0.0%	0.1%	0.2%	0.3%	0.5%
Alternative Data Usage 2	17	4.5%	2.7%	1.7%	2.2%	4.0%	5.9%	11.8%
Alternative Data Usage 3	17	10.8%	4.3%	2.8%	7.9%	11.2%	13.3%	18.5%

19 List of keywords

“Source: iSentia” “Source: ListenFirst” “Source: Meltwater” “Source: Metricle” “Source: Sharablee” “Source: Social Alpha” “Source: Social Market Analytics” “Source: SumZero” “Source: TheySay” “Source: Trackur” “Source: TrustedInsight” “Source: TYR” “Source: Zhiwei” “Source: 411.info” “Source: 7Park” “Source: 90 West Data” “Source: 90 West” “Source: AADC” “Source: AAER dataset” “Source: Accern” “Source: AccountScore” “Source: Acuris” “Source: AddThis” “Source: AddThis.com” “Source: Affinity Solutions” “Source: Affinity” “Source: Affirm” “Source: AHA” “Source: Airports Council International” “Source: Alexa” “Source: Alexa.com” “Source: Alexandria” “Source: AllTheRooms” “Source: Almax Information Systems” “Source: AlphaFlow” “Source: Alphamatician ” “Source: Alphamatician” “Source: Alphasense” “Source: Amareos” “Source: Amenity Analytics” “Source: American Chemistry Council” “Source: American Trucking Associations” “Source: App Annie” “Source: Arab Air Carrier Organization” “Source: ARC” “Source: ARM Insight” “Source: Ascend Worldwide Limited” “Source: Asset Macro” “Source: Associated Press” “Source: Audit Analytics” “Source: Australia Data” “Source: Australian Antarctic Data Centre” “Source: Australian Associated Press” “Source: BayStreet Research” “Source: Beijing Chuang Yi Fang Technology” “Source: Beijing UC Science & Technology” “Source: BillGuard” “Source: Bitly” “Source: BizQualify” “Source: Black Box” “Source: Bloomberg Tesla Tracker” “Source: Borrell” “Source: Boxoffice Media LLC” “Source: Boxoffice Media” “Source: Boxoffice” “Source: Brain Company” “Source: BrandLoyalties” “Source: BrandWatch” “Source: Brave New Coin” “Source: Brazil” “Source: Bridg” “Source: Broughton Capital” “Source: Buddy” “Source: Buildfax” “Source: Business Intelligence Advisors” “Source: Business Monitor International” “Source: Canada Open Data” “Source: CB Richard Ellis Inc.” “Source: CDU TEK: Central Dispatching Department of Fuel Energy Complex of Russia” “Source: CEIC Data” “Source: CESSDA” “Source: Chain Store Guide Information Services” “Source: ChemOrbis” “Source: CHINA CCM” “Source: China Government data” “Source: China Money Network” “Source: China National Chemical Information Center” “Source: China Real Estate Information Corporation” “Source: CIA Data” “Source: Cignifi” “Source: City Data” “Source: Civic Science” “Source:

Cognical" "Source: CogniSent" "Source: Comlinkdata" "Source: Companies House UK" "Source: CompStak" "Source: Comscore Inc." "Source: Comscore" "Source: Consumer Edge" "Source: Cooltrader" "Source: CoreLogic" "Source: CQG" "Source: Crain Communications Inc." "Source: Credit Ease" "Source: CreditRiskMonitor" "Source: CRIC" "Source: Crimson Hexagon" "Source: Cropnosis" "Source: CropProphet" "Source: CU Steel" "Source: Cuebiq" "Source: Data Catalogs" "Source: Data Norfolk" "Source: Data Simply" "Source: Data.NSW" "Source: DataFerrett" "Source: Datalogix" "Source: Dataminr" "Source: Datamyne" "Source: Dataprovider" "Source: Dataprovider.com" "Source: DataPulse" "Source: Datarama" "Source: DataSift" "Source: DataStreamX" "Source: DataTrek" "Source: DataWeave" "Source: DataYes" "Source: Delphia" "Source: Demyst" "Source: DemystData" "Source: Descartes Labs" "Source: Discern" "Source: Doane Advisory Service" "Source: Dodge" "Source: Drawbridge" "Source: Drewry Shipping Consultants Ltd" "Source: Drillinginfo" "Source: DTCC" "Source: Dun & Bradstreet" "Source: Earnest" "Source: EcommerceDB" "Source: EconData" "Source: Edgar Online" "Source: Edison trends" "Source: Edmunds" "Source: EIDO" "Source: EIDOSearch" "Source: Eilers & Krejcik Gaming" "Source: Eilers" "Source: Elevate" "Source: Enigma" "Source: Enova" "Source: Entgroup" "Source: Envestnet Yodlee Data Analytics" "Source: Envestnet Yodlee" "Source: Envestnet" "Source: Environment Agency UK" "Source: EODData" "Source: EPFR" "Source: Epsilon" "Source: eSignal" "Source: Estimize" "Source: Eurekahedge" "Source: Euromonitor International" "Source: European Union Open data" "Source: EventVestor" "Source: Everest Group" "Source: Exante Data" "Source: Exerica" "Source: Experian Footfall" "Source: Experian Micro Electronics" "Source: Experian" "Source: Fable Data" "Source: Fable" "Source: Facebook" "Source: Facteus" "Source: FactorTrust" "Source: FactSet Revere" "Source: FairLoan" "Source: Fashionbi" "Source: FastBooking" "Source: Federal Reserve Bank of St. Louis" "Source: FedStats" "Source: FHS Swiss Watch Data" "Source: FICO" "Source: Finadium" "Source: Finweavers" "Source: First Data Merchant Services Corporation" "Source: First Data Spend" "Source: First Data SpendTrend" "Source: First Data" "Source: FirstAccess" "Source: FN Arena" "Source: FNGO" "Source: France Government Data" "Source: FRED" "Source: FTR Freight Transport Research Associates" "Source: Funding Circle" "Source: GB Office for National Statis-

tics” “Source: GDELT Project” “Source: Genscape” “Source: GeoQuant” “Source: Germany Government Data” “Source: Gildata” “Source: Glassdoor” “Source: Global Health Observatory” “Source: Global Open Data Index” “Source: GovSpend” “Source: Grandata” “Source: Gro Intelligence” “Source: Happy Mango” “Source: Haver Analytics” “Source: Health Forum” “Source: Heckyl” “Source: HFR” “Source: Hillside Partners” “Source: Hillside” “Source: Hong Kong Data” “Source: IANA” “Source: ICI” “Source: ICRA Services” “Source: IhsMarkit” “Source: iiMedia Research” “Source: IMF DATA” “Source: IMS Quintiles” “Source: Index Marketing Solutions Limited” “Source: India Open Government Data” “Source: Inferess” “Source: Informa Financial” “Source: Informa Financial-Intelligence” “Source: InfoTEK Publishing House” “Source: InfoTrie” “Source: Innovata” “Source: Innovate UK” “Source: Inside Mortgage Finance Publication” “Source: Insights Data Solutions” “Source: InSpectrum” “Source: Instagram” “Source: Intelius” “Source: Interconnect Analytics” “Source: Intermodal Association of North America” “Source: International Data Corporation Inc.” “Source: International Labour Organization” “Source: Internet Truckstop” “Source: Intrinio” “Source: Invyo” “Source: iResearch” “Source: iSentium” “Source: ISSB Ltd” “Source: IT Finance” “Source: Italy” “Source: Jackson County GIS” “Source: Japan Government Data” “Source: Jettrack.io” “Source: Jiguang” “Source: Jumpshot” “Source: JustData” “Source: JWN Energy” “Source: Kabbage” “Source: KD Interactive” “Source: Knowsis” “Source: Korea” “Source: Kpler” “Source: Kreditech” “Source: Kyber Data Science” “Source: Kyber” “Source: Lenddo” “Source: LendingClub” “Source: LendUp” “Source: Lexalytics” “Source: LexisNexis” “Source: LikeFolio” “Source: LIMRA” “Source: LinkUp” “Source: Liquor and Cannabis Board data portal” “Source: LL Global Inc.” “Source: Lota Data” “Source: Lota” “Source: Lucena” “Source: M&A Portal” “Source: MAC Data” “Source: Magna Global Research” “Source: Manfredi & Associates” “Source: Manheim” “Source: MariData” “Source: MarketCheck” “Source: MarketPsych” “Source: Marketscout Corporation” “Source: Markit Securities Finance” “Source: MASSIVE Data Heights” “Source: MasterCard Advisors” “Source: MatterMark” “Source: MedMine” “Source: Mexico” “Source: MIDiA Research” “Source: MIDiA” “Source: Millennium Research” “Source: MixRank” “Source: Money Dashboard” “Source: MoneySuperMarket” “Source: Moodyâ€™s Data Alliance” “Source: Neotrade Analytics” “Source:

Netspend" "Source: New Zealand" "Source: Newscred" "Source: Newswhip" "Source: Nexant" "Source: NIC" "Source: NIELSEN Data" "Source: Norway" "Source: NYC Opendata" "Source: OECD Data" "Source: OECD.Stat" "Source: Off Highway Research Limited" "Source: Omega Point " "Source: Omney Data" "Source: One Click Retail" "Source: OpenSignal" "Source: OPPNADATA" "Source: Optimum Complexity" "Source: Orbital Insights" "Source: Oregon state data" "Source: Ovum Ltd Us Branch" "Source: P&I Research" "Source: Panjiva" "Source: Panvista Analytics" "Source: PAYDEX" "Source: PayLead" "Source: Photon Commerce" "Source: Placemeter" "Source: Pluribus Labs" "Source: Predata" "Source: Predict HQ" "Source: Premise" "Source: PriceStats" "Source: Pro Publica" "Source: Prosper Insights & Analytics" "Source: Prosper Insights" "Source: Prosper Marketplace" "Source: Qlik" "Source: Quandl" "Source: Quant Connect" "Source: Quantcube" "Source: Queensland Government Data" "Source: Quest Offshore" "Source: QueXopa" "Source: RandomWalk" "Source: RavenPack" "Source: RCMD" "Source: Re analytics" "Source: Real Capital Analytics" "Source: Reanalytics" "Source: Redbook Research Inc." "Source: RedTech" "Source: REIS" "Source: RelateTheNews" "Source: RelationshipScience" "Source: RepRisk" "Source: Repustate" "Source: Return Path" "Source: RevolutionCredit" "Source: Rigdata" "Source: RigLogix" "Source: Root Metrics" "Source: RootMetrics" "Source: RS Metrics" "Source: RunningAlpha" "Source: Russia" "Source: RxData.net" "Source: Rystad Energy" "Source: Sandalwood" "Source: Scoop" "Source: Scutify" "Source: Second Measure" "Source: Seer Aerospace" "Source: Selerity" "Source: SEMI" "Source: Semiconductor Equipment & Materials International" "Source: Semlab" "Source: Sensor Tower" "Source: Sentifi" "Source: Sentiment Trader" "Source: Sequentum" "Source: SESAMm" "Source: Shanghai Metals Market" "Source: ShareIQ" "Source: ShareThis" "Source: Shoppertrak Rct Corporation" "Source: ShopperTrak" "Source: Sigmai" "Source: Signal.co" "Source: Singapore Data" "Source: SJ Consulting Group Inc." "Source: Sky Watch" "Source: Skydeo" "Source: Slice Intelligence" "Source: SmarterWorks" "Source: Smith Travel" "Source: SNL Kagan" "Source: Social Market" "Source: Social Media" "Source: Social Network" "Source: Socrata" "Source: South Africa" "Source: Spain" "Source: Standard Media Index" "Source: Stax" "Source: Steel Orbis" "Source: StockTwits" "Source: STR" "Source: SuperData" "Source: Superfly insights" "Source:

SuperFly" "Source: Sustainalytics" "Source: Sweden" "Source: Taiwan" "Source: Tala" "Source: TDn2k" "Source: Tecnon Orbichem" "Source: Teragence" "Source: The BIS International Financial Statistics" "Source: The Fertilizer Institute" "Source: Think Finance" "Source: ThinkNum" "Source: Tick Data" "Source: Tipigo" "Source: TMT Analysis" "Source: Tobacco Merchants Assoc Inc." "Source: Toyo Keizai" "Source: Trade and Tariff Data" "Source: Tradesparq" "Source: Trading Economics" "Source: Transport Topics Publishing Group" "Source: Tribe Dynamics" "Source: Triton Research" "Source: TrustData" "Source: TVeyes" "Source: Twitter" "Source: TXN" "Source: Uber Media" "Source: UBS Evidence Lab" "Source: UK Healthcare Data" "Source: UK" "Source: UN Data" "Source: UNICEF Data" "Source: UnionPay" "Source: United Nations" "Source: Unmetric" "Source: US Census Bureau" "Source: US Census" "Source: US Government Data" "Source: US Healthcare Data" "Source: VantageScore" "Source: Veronis Suhler Stevenson" "Source: VesselsValue" "Source: Vestdata" "Source: Victorian Government open data" "Source: Vigilant" "Source: VisualDNA" "Source: Wards Automotive Group" "Source: WDZJ.com" "Source: Webhose.io" "Source: Wind" "Source: World Bank" "Source: World Bureau of Metal Statistics" "Source: World Container Index" "Source: XIQ" "Source: Yipit" "Source: YipitData" "Source: Yodlee" "Source: Zephyr" "Source: ZestFinance"

References

Antweiler, Werner, and Murray Frank, 2004, Is all that talk just noise? the information content of internet stock message boards, *Journal of Finance* 52, 1259–1294.

Bartov, Eli, Lucile Faurel, and Partha Mohanram, 2018, Can Twitter help predict firm-level earnings and stock returns?, *The Accounting Review* 93, 25–57.

Bollen, Johan, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of Computational Science* 2, 1–8.

Chen, Hailang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, *Review of Financial Studies* 27, 1367–1403.

Choi, Hyunyoung, and Hal Varian, 2012, Predicting the present with google trends, *Economic Record* 88, 2–9.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In search of attention, *The Journal of Finance* 66, 1461–1499.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock, 2012, Investor information demand: Evidence from google searches around earnings announcements, *Journal of Accounting Research* 50, 1001–1040.

Froot, Kenneth, Namho Kang, Gideon Ozik, and Ronnie Sadka, 2017, What do measures of real-time corporate sales say about earnings surprises and post-announcement returns?, *Journal of Financial Economics* 125, 143–162.

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.

Goetzmann, William N., Dasol Kim, Alok Kumar, and Qin Wang, 2014, Weather-Induced Mood, Institutional Investors, and Stock Returns, *The Review of Financial Studies* 28, 73–111.

Green, T. Clifton, Ruoyan Huang, Quan Wen, and Dexin Zhou, 2019, Crowdsourced employer reviews and stock returns, *Journal of Financial Economics* 134, 236–251.

Gu, Chen, and Alexander Kurov, 2020, Informational role of social media: Evidence from twitter sentiment, *Journal of Banking and Finance* 121, 1059–1069.

Hirschey, Mark, Vernon J. Richardson, and Susan Scholz, 2000, Stock-price effects of internet buy-sell recommendations: The motley fool case, *Financial Review* 35, 147–174.

Hirshleifer, David, and Tyler Shumway, 2003, Good day sunshine: Stock returns and the weather, *The Journal of Finance* 58, 1009–1032.

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

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

Katona, Zsolt, Marcus Painter, Panos N. Patatoukas, and Jean Zeng, 2021, On the capital market consequences of alternative data: Evidence from outer space, Working paper Saint Louis University.

Kelley, Eric K., and Paul C. Tetlock, 2013, How wise are crowds? insights from retail orders and stock returns, *The Journal of Finance* 68, 1229–1265.

Leung, Woon, Gabriel Wong, and Woon Wong, 2019, Social-media sentiment, portfolio complexity, and stock returns, Working paper University of Edinburgh.

Mukherjee, Abhiroop, George Panayotov, and Janghoon Shon, 2021, Eye in the sky: Private satellites and government macro data, *Journal of Financial Economics* 141, 234–254.

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

Tumarkin, Robert, and Robert F. Whitelaw, 2001, News or noise? internet postings and stock prices, *Financial Analysts Journal* 57, 41–51.

Umar, Tarik, 2022, Complexity aversion when seeking alpha, *Journal of Accounting and Economics* 73, 1014–1077.

Wu, Lynn, and Erik Brynjolfsson, 2015, *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales* . chap. 3, pp. 89–118 (University of Chicago Press).

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