Online Appendix for

Does Alternative Data Improve Financial Forecasting? The Horizon Effect

(not intended for publication)

July 20, 2021

Contents

1 Dividing Forecasting Tasks

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

1.1 Duplication of fixed costs of information production

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

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

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

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

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

Given the signals reported by the agents, the analyst forms her forecasts optimally, as in the baseline model. Thus, proceeding as in the baseline model, the analyst's objective function at date 0 is to choose $\{z_{st}^{**}, z_{lt}^{**}, \omega_{st}^{*}, \omega_{lt}^{*}\}$ solving

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Agency costs.

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

1.3 Complementarity and imperfect communication

The analysis in Section 1.1 implicitly assumes that the tasks of obtaining information about the common component and the unique component of the long-term earnings can be separated. A more plausible assumption is that achieving the first task is necessary to achieve the second one. Intuitively, one cannot obtain a signal about the unique component of the long-term earnings without first filtering out the common component from information about the long-term earnings. The reverse is not true because information about the common component can be obtained by just focusing on information relevant for forecasting the short-term earnings. Thus, it is natural to see the tasks of obtaining signals about the common and the unique components in the firm's earnings as being "ordered." Achieving the first task (obtaining a signal about the short-term earnings, i.e., the common component of firms' earnings) is a necessary prerequisite for achieving the second one (obtaining a signal about the unique component of the long-term earnings). This ordering creates a form of complementarity between the two tasks: the second task yields a signal only if the first one has been completed.

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

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

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

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

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

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

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

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

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

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

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

2 Shock on ψ_{st}

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

First, it is direct from Proposition 1 that

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

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

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

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

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

3 EPS to Net Income Forecast Conversion

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

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

4 Why not Use Long-Term Growth Forecasts?

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

5 Forecast Informativeness with Biased Analysts

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

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

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

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

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

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

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

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

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

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

$$R_{ih}^2=\widehat{k}_1^2\frac{Var(f_{hi})}{Var(\theta_{hi})}=\widehat{k}_1R_{\theta f^*}^2 \qquad (10)$$

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

6 Analysts' Forecasting Activity and Recommendations

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

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

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

Table A1: Social Media Data and Analysts' Forecasting Activity

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

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

Table A2: Social Media Data and Analysts' Recommendations

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

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

7 Actual Messages vs. Hypothetical Messages

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

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

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

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

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

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

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

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

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

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

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

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

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

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes all available analyst-day-horizon observations between 2005 and 2017. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). w'_i is the mean of $\overline{w_j}$ across the firms covered by the analyst. $\overline{w_j}$ is the mean of $w_{j,t}$ for all t after a message is observed for the first time about j. Other control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. In columns (2), (3), analyst and date fixed effects are interacted with h^*. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)		
Data Exposure Proxy:	# Hypothetical Messages		
OLS:	(1)	(2)	(3)
$h^*\times$ Data Exposure	-1.15***	-1.03***	-1.05***
	(-3.72)	(-4.40)	(-5.03)
Data Exposure	0.05	-0.39	-0.4
	(0.28)	(-1.57)	(-1.60)
$h^*\times$ w'_i	3.54***	0.33	-0.21
	(2.37)	(0.20)	(-0.13)
w'_i	2.79***	3.46***	1.3
	(2.42)	(2.62)	(0.93)
h^*	-16.77***		
	(-32.69)		
Analyst FE	Yes		
Date FE	Yes		
Analyst FE (interacted)		Yes	Yes
Date FE (interacted)		Yes	Yes
Controls			Yes
N	30,959,276	30,105,551	27,860,424

8 Do Our Measures Correlate with News from Standard Sources?

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

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

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

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

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

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

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

This table presents OLS estimates of the sensitivity of different measures of social media data exposure to news arrival from standard sources. Estimations are made at the firm-day level. The sample includes all U.S. firms that have been discussed at least once on StockTwits between 2009 and 2017, and that are covered by at least one analyst. In columns (1) to (3), *#Watchlist* is the number of StockTwits users having the firm in their watchlist on day t. In columns (4) to (6), *#Hypothetical Messages* is the number of hypothetical messages posted about the firm from $t-30$ to $t-1$. In columns (7) to (9), *#Messages* is the number of actual messages posted about the firm from $t-30$ to $t-1$. #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)
#News$_t$	-4.66			-2.82			5.67***		
	(-0.59)			(-0.82)			(2.97)		
#News$_{t-1}$		-3.98			-1.95			9.11***	
		(-0.51)			(-0.59)			(4.65)	
#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 #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 $\#News_{t\rightarrow T}$ 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 #News$_t$	1.35			-0.44			5.14***		
	(0.80)			(-0.44)			(6.38)		
Mkt Resp. to #News$_{t-1}$		1.60			-0.32			6.56***	
		(0.95)			(-0.33)			(7.74)	
Mkt Resp. to #News$_{t-30\rightarrow t-1}$			-0.30			-0.67			4.91***
			(-0.26)			(-0.98)			(10.31)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902	18,568,413	18,566,389	16,996,902

9 Robustness Table II

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

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

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

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

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

Table A6: Robustness: Forecast Informativeness by Horizon

This table presents OLS estimates of time trend in analysts' forecasts' informativeness by sub-samples including observations in different annual forecasting horizon ranges. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. h is the forecasting horizon, measured as the number of days between the forecasting date and the date of actual earnings release, divided by 365. Year Trend is a variable that takes the value of zero for the period 1983-1992 and increments by one every subsequent year, divided by 25 so that the regression coefficient can be interpreted as the cumulative increment in R^2 over the 1993-2017 period. Variable definitions are in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

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

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

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

10 Robustness Table III

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

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

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

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

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

Table A7: The Slope of the Term Structure

This table presents OLS estimates of time trend in the slope of the term structure of forecasts' informativeness. The dependent variable is the slope of the term structure. This slope measures the change in R^2 (in percentage points) when horizon increases by one year. A negative slope indicates that forecasts' informativeness (R^2) decreases with horizon. In column (1), the slope is calculated every year by regressing the average of R^2 by horizon on the horizon h (i.e., the number of days between the forecasting date and the date of actual earnings release, divided by 365). In columns (2) and (3), the slope is calculated every year by 2-digit SIC industry by regressing the average of R^2 by horizon and industry on h. In columns (4) and (5), the slope is calculated every year by analyst by regressing the average of R^2 by horizon and analyst on h. Year Trend is a variable that takes the value of zero for the period 1983-1992 and increments by one every subsequent year divided by 25 so that the regression coefficient can directly be interpreted as the cumulative change in slope over the 1993-2017 period. Variable definitions are in the Appendix II. t-statistics in parentheses are based on standard errors clustered by year. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

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

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

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

11 Robustness Table VI

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

Table A8: Robustness: Controlling for Trading Volume

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes all available analyst-day-horizon observations between 2005 and 2017. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h-1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). Trading volume is the total number of shares traded from $t-30$ to $t-1$, measured first by firm and then averaged across the firms covered by analysts. Other control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Data Exposure:		*#Watchlist*			*#Hypothetical Messages*	
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^* \times$ Data Exposure	-1.09***	-0.86***	-1.00***	-1.01***	-1.06***	-1.13***
	(-3.23)	(-3.17)	(-3.74)	(-3.88)	(-4.84)	(-5.32)
Data Exposure	0.16	-0.17	-0.3	0.38*	-0.14	-0.25
	(0.66)	(-0.68)	(-1.15)	(1.62)	(-0.62)	(-1.03)
$h^* \times$ Trading Volume	1.13***	0.62***	0.57***	1.18***	0.71***	0.66***
	(6.56)	(3.28)	(2.67)	(6.82)	(3.76)	(3.17)
Trading Volume	-0.4	-0.12	-1.23***	-0.43	-0.12	-1.23***
	(-1.29)	(-0.49)	(-3.80)	(-1.39)	(-0.48)	(-3.83)
h^*	-17.62***			-17.59***		
	(-31.69)			(-30.94)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,276	30,105,551	27,860,424	30,959,276	30,105,551	27,860,424

Table A9: Robustness: Analysts With Stable Coverage

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with stable coverage only. Coverage is stable if the level of similarity between the portfolio of firms covered in the current year and that of the previous year is greater than 90%. Similarity is defined as the number of common firms between the portfolio covered in the current year and the one covered the year before, scaled by the square root of the product of the number of firms in each portfolio. The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Data Exposure:	*#Watchlist*			*#Hypothetical Messages*		
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^*\times$ Data Exposure	-0.46	-0.50**	-0.69***	-0.29	-0.71***	-0.85***
	(-1.49)	(-2.02)	(-2.60)	(-1.26)	(-3.46)	(-3.82)
Data Exposure	0.32	0.04	-0.15	0.48*	0.00	-0.16
	(1.25)	(0.15)	(-0.52)	(1.68)	(0.01)	(-0.64)
h^*	-16.35***			-16.34***		
	(-36.86)			(-35.24)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	14,552,288	13,773,488	12,683,367	14,552,288	13,773,488	12,683,367

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

This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts with non-missing long-term forecasts. An analyst has non-missing long-term forecasts if there is at least one non-missing $R^2_{i,t,h}$ for $h \geq 3$ over the sample period (2005-2017). The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist (*#Watchlist*), or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$ (*#Hypothetical Messages*). h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h - 1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Data Exposure:		*#Watchlist*			*#Hypothetical Messages*	
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^* \times$ Data Exposure	-1.40***	-1.07***	-1.25***	-1.10***	-1.19***	-1.27***
	(-4.17)	(-3.34)	(-4.13)	(-4.59)	(-5.96)	(-7.42)
Data Exposure	-0.12	-0.29	-0.48	0.16	-0.26	-0.39
	(-0.54)	(-1.00)	(-1.49)	(0.71)	(-1.08)	(-1.58)
h^*	-15.33***			-15.24***		
	(-41.16)			(-38.40)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	13,782,999	13,019,477	12,153,633	13,782,999	13,019,477	12,153,633

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

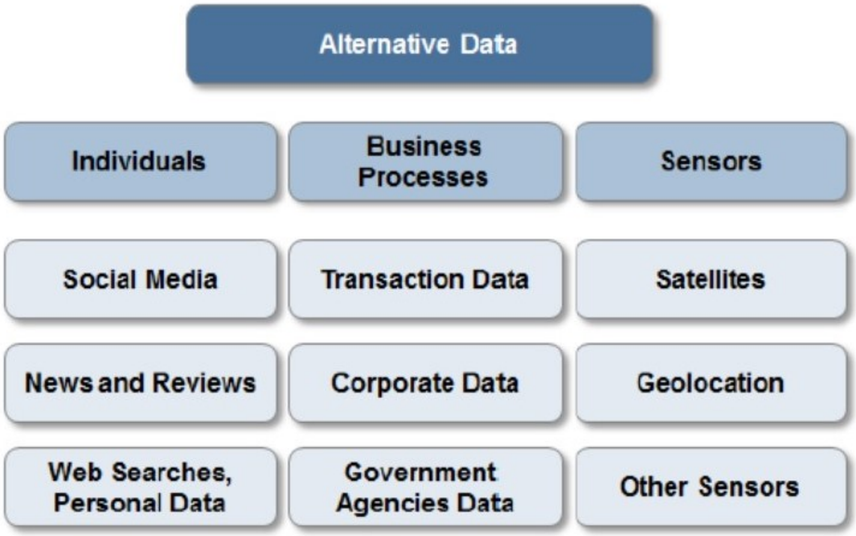
This table presents OLS estimates of the sensitivity of the informativeness of analysts' forecasts (R^2) to social media data generated on StockTwits (eq.(18)). The sample includes analyst-day-horizon observations between 2005 and 2017 for analysts covering firms with non-missing long-term forecasts. A firm has non-missing long-term forecasts if it has at least one non-missing forecast for $h \geq 3$ over the sample period (2005-2017). The dependent variable is R^2, which measures the informativeness of the forecasts made by an analyst on a given day for a given horizon. Data Exposure is a variable capturing the exposure to data generated on StockTwits, measured first by firm and then averaged across the firms covered by analysts at time $t-1$, where t is the date at which we measure forecast informativeness. Data Exposure is set to zero prior to StockTwits' introduction in 2009, and normalized by its in-sample standard deviation. Data Exposure is based on the average number of users that have the firms covered by the analyst in their watchlist, or the number of hypothetical messages posted about those firms from $t-30$ to $t-1$. h is the forecasting horizon, measured as the number of days between t and the date of actual earnings release, divided by 365. h^* is the forecasting horizon centered at 1 ($h^* = h-1$) so that the regression coefficient on the baseline variable Data Exposure can be interpreted as the unconditional effect on R^2 at the one-year horizon (rather than zero). In columns (2), (3), (5), and (6), analyst and date fixed effects are interacted with h^*. Control variables include firms' cash flow to assets, cash to assets, debt to assets, Tobin's Q, the log of total assets, and the log of age, calculated using the last available financials and averaged by analyst at time $t-1$. Detailed variable definitions are provided in Appendix II. t-statistics in parentheses are based on standard errors clustered by forecasted fiscal period. Symbols ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Dep. variable:	Forecast informativeness (R^2)					
Data Exposure:		*#Watchlist*			*#Hypothetical Messages*	
OLS:	(1)	(2)	(3)	(4)	(5)	(6)
$h^* \times$ Data Exposure	-0.86***	-0.78***	-0.96***	-0.69***	-0.94***	-1.05***
	(-2.59)	(-3.06)	(-3.72)	(-2.75)	(-4.54)	(-5.03)
Data Exposure	0.13	-0.17	-0.35	0.34	-0.14	-0.32
	(0.50)	(-0.64)	(-1.29)	(1.42)	(-0.57)	(-1.30)
h^*	-16.66***			-16.62***		
	(-33.85)			(-32.13)		
Analyst FE	Yes			Yes		
Date FE	Yes			Yes		
Analyst FE (interacted)		Yes	Yes		Yes	Yes
Date FE (interacted)		Yes	Yes		Yes	Yes
Controls			Yes			Yes
N	30,959,281	30,105,556	27,860,429	30,959,281	30,105,556	27,860,429

12 Alternative Data: Definition and Classification

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

Figure 1: Classification of big/alternative data sources



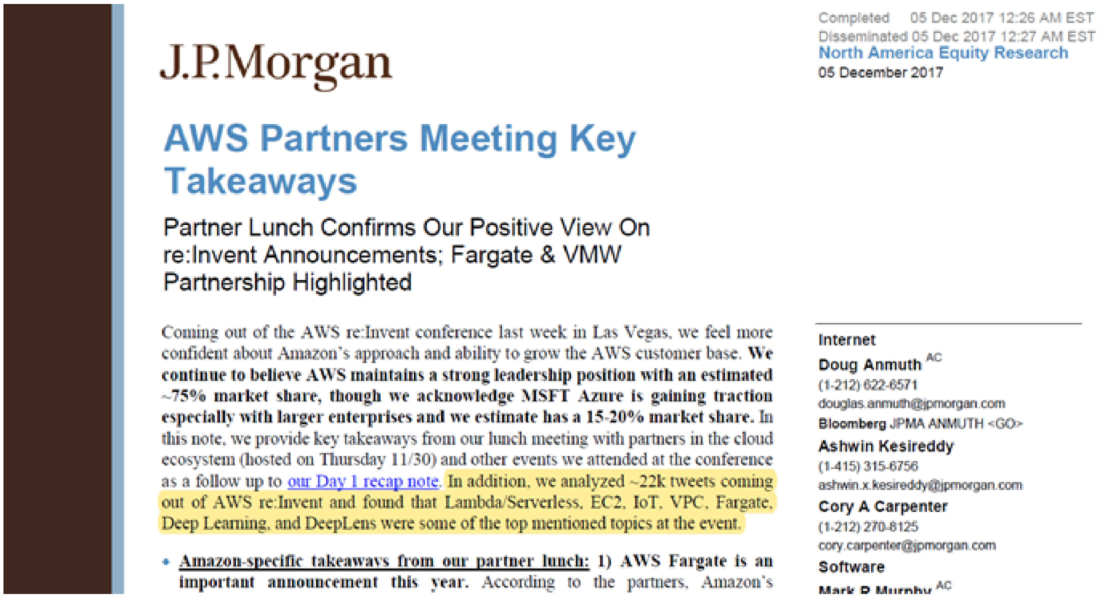
Alternative Data

Individuals	Business Processes	Sensors
Social Media	Transaction Data	Satellites
News and Reviews	Corporate Data	Geolocation
Web Searches, Personal Data	Government Agencies Data	Other Sensors

Source: J.P. Morgan QDS

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

13 Example of Analysts Using Social Media Data



J.P.Morgan

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

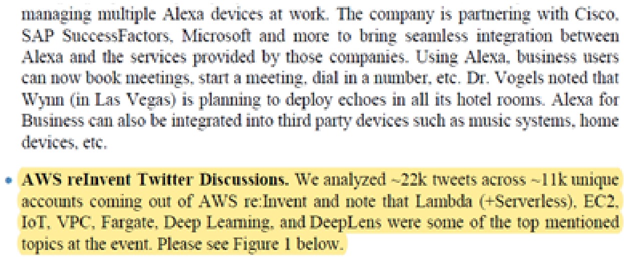
AWS Partners Meeting Key Takeaways

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

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

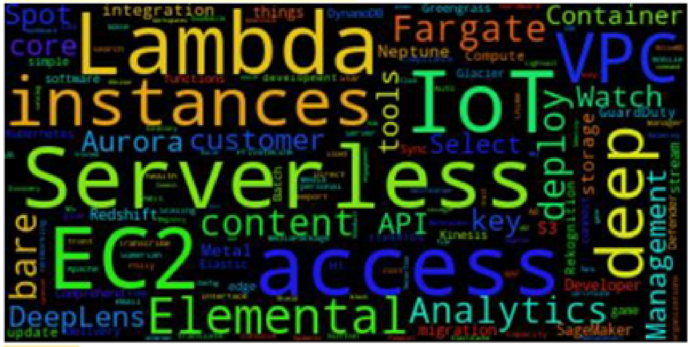
Internet
Doug Anmuth AC
(1-212) 622-6571
douglas.anmuth@jpmorgan.com
Bloomberg JPMA ANMUTH <GO>
Ashwin Kesireddy
(1-415) 315-6756
ashwin.x.kesireddy@jpmorgan.com
Cory A Carpenter
(1-212) 270-8125
cory.carpenter@jpmorgan.com
Software
Mark R Murphy AC



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.