Alternative Data and Forecasting Horizon: Does More Data Mean Better Forecasts?

Thierry Foucault (HEC Paris and CEPR)

Based on joint work with Olivier Dessaint (INSEAD) and Laurent Frésard (USI Lugano) (available at https://dx.doi.org/10.2139/ssrn.3702411).

Part of the research discussed in this presentation has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No 101018214)

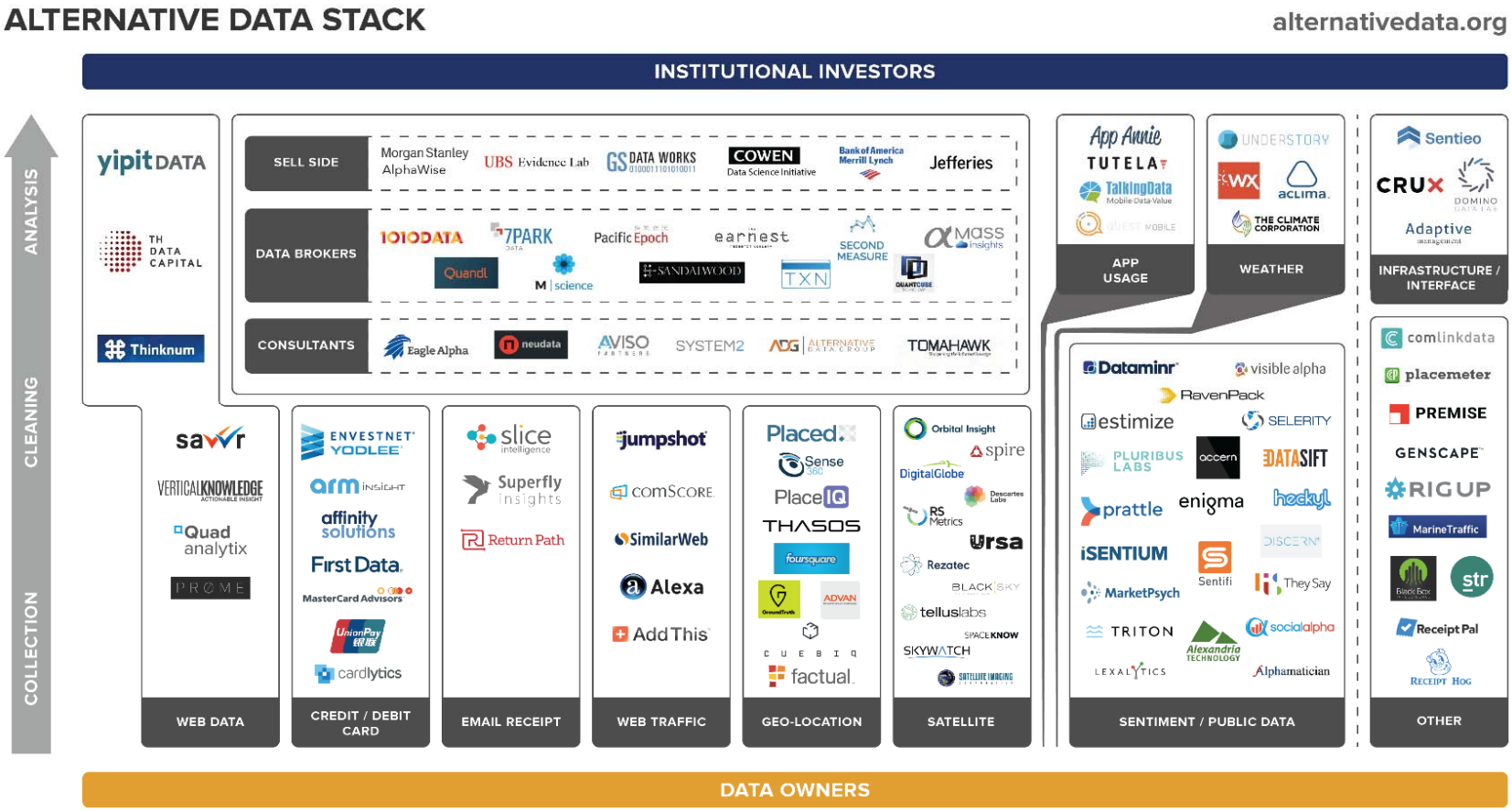
AI and Finance

- **Artificial Intelligence in Finance:**
  - New forecasting techniques ("machine learning")

    AND

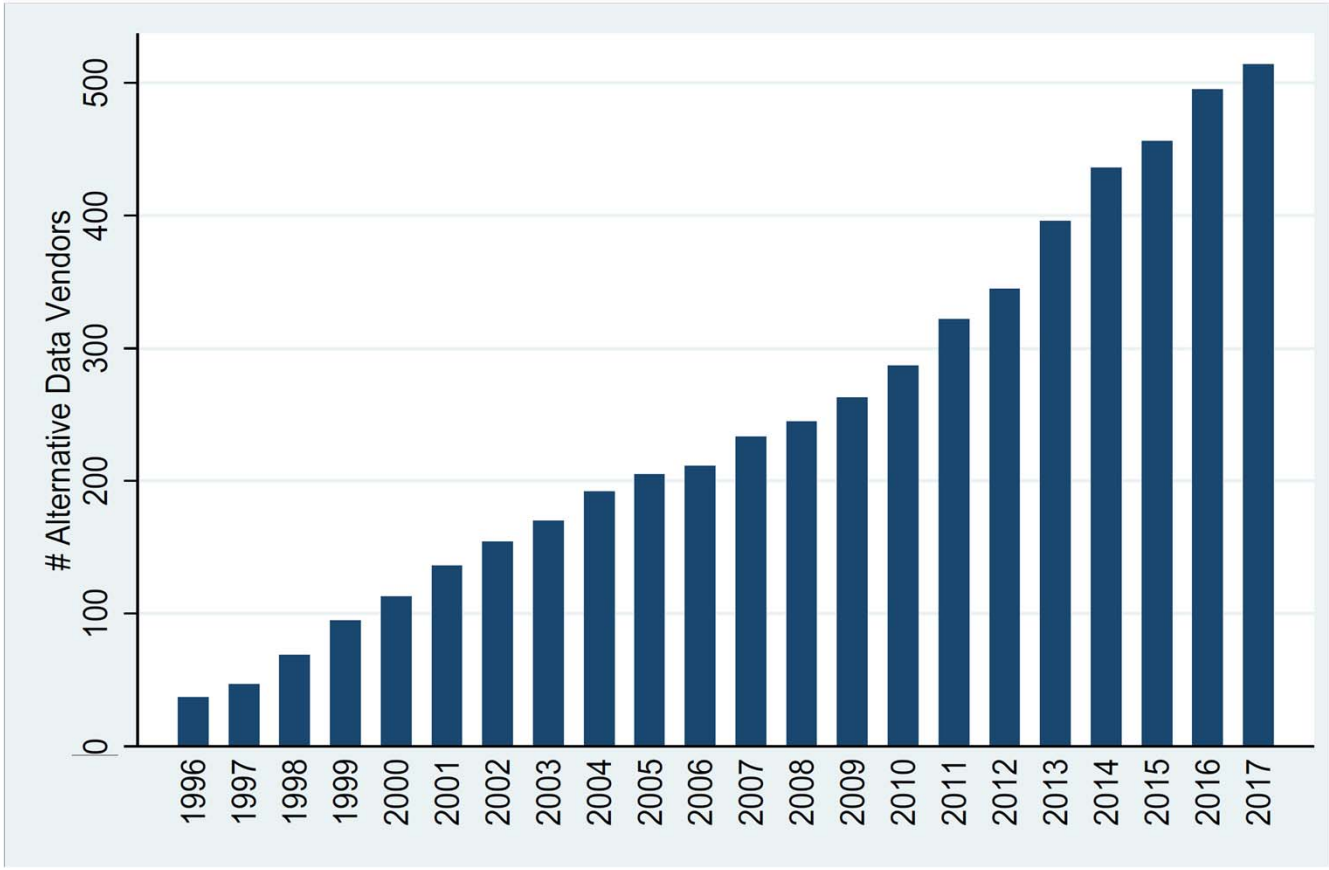
  - New data: "Alternative Data" (social media, credit card and point of sale, employee satisfaction, online searches, satellite images, mobile phones etc.)
- Does **data abundance** improve the quality of financial forecasts (forecasts about cash-flows and returns)?

Data Abundance



Source: Alternativedata.org (https://alternativedata.org/)

Alternative Data providers



Source: Dessaint, Foucault, Frésard (2022). Calculations using JP Morgan Alternative Data Handbook and the Internet Archive Wayback machine.

Is there Information in Alt. Data

- Yes….BUT mostly short-term information.

- Evidence from 26 academic papers using alternative data (social media, employer ratings, satellite, google searches etc) for predictability of sales/earnings or returns.

  None finds predictability beyond 1 year!

- Alternative Data = "Short-Term Oriented" Data

Questions

- Does greater availability of **short-term oriented** data improve financial forecasting?
  - Standard economic reasoning on this question: More data reduces the cost of obtaining information $\Rightarrow$ Agents acquire more information and their forecasts improve.
  - But financial decisions (e.g., asset valuation) often require forecasting short-term **and** long-term outcomes (e.g., earnings)
- What happens to the quality ("informativeness") of long-term (LT) forecasts when the cost of accessing short-term (ST) information decreases? Does the quality of both ST and LT forecasts improve? Not necessarily…

Main Message

- A reduction in the cost of accessing short-term oriented data can reduce the quality of LT forecasts
  - **Theory:** Reducing the cost of obtaining short-term information induces forecasters to focus even more on the short-term (low cost) at the expense of their attention to long-term (high cost).
  - **Supporting evidence using** sell-side equity analysts forecasts

Roadmap

Theory and predictions

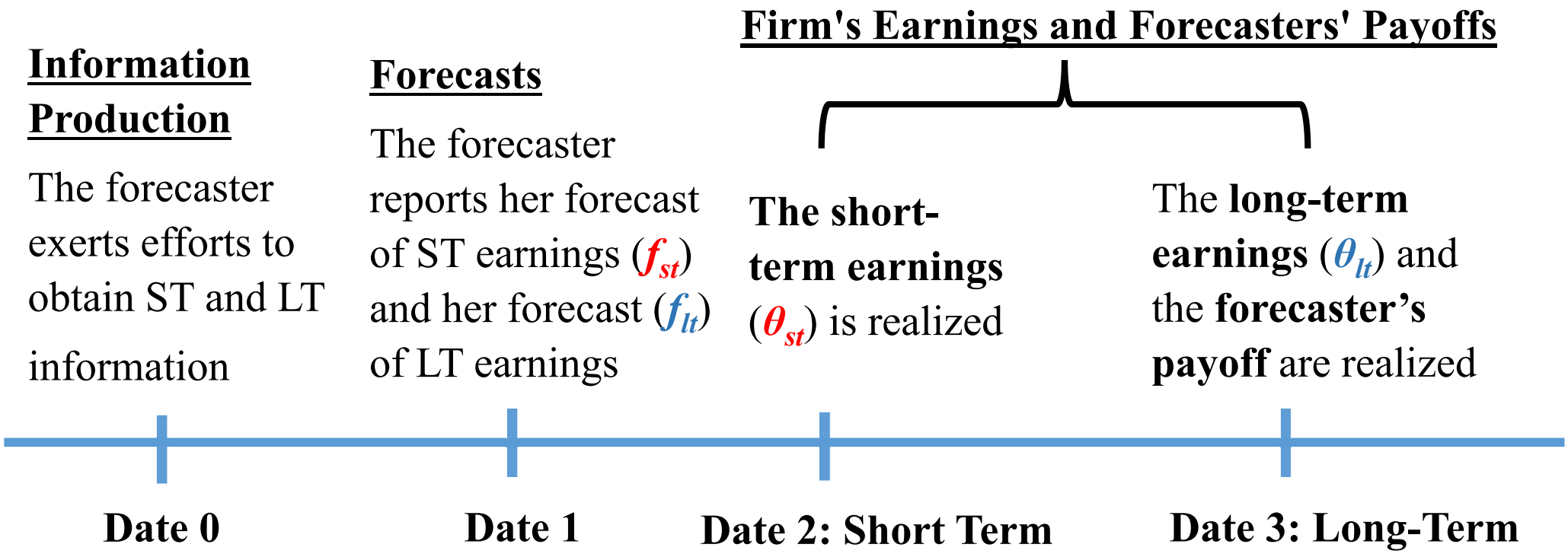
Measuring analysts' forecasts informativeness

Long-run trend in analysts forecasts' informativeness and alternative data

Main test: new social media data

Take-away and implications

Theory (Sketch)



Key Assumptions

1. Forecasting short-term and long-term earnings are related but **distinct tasks**
   - Earnings are correlated over time but not perfectly correlated
2. **Effort pays:** The more effort allocated to collect information relevant to predict earnings at a given horizon, the more accurate are signals received to forecast these earnings.
3. **Effort is costly:** The more effort allocated to collect information for predicting earnings at one horizon, the higher the cost (cognitive, time etc.).
   - See Harford et al. (2018) and Hirshleifer et al. (2019) for evidence on security analysts
4. **Multitasking is costly:** More effort allocated to one task makes the effort allocated to the other one more costly
5. **Availability of Short-Term Oriented data reduces the cost of obtaining short-term information**

Forecaster's Objective

- The forecaster chooses her efforts to obtain ST and LT information at date 0 and her forecasts at date 1 to **minimize the weighted sum of her expected squared forecasting errors plus the total cost of obtaing ST and LT information.**

- **Trade-off:** More efforts
  - Reduce expected squared forecasting errors
  - Increase cost of obtaining information

- **Optimal solution to this trade-off:** Chooses efforts so that marginal benefit= marginal cost of efforts.

Implications (Predictions)

- The availability of Short-Term Oriented data reduces the cost of obtaining ST information and induces the forecaster to optimally
  - Increase her effort to obtain short-term Information
  - Reduce her effort to obtain long-term information due to multitasking costs

- $\Rightarrow$ The forecaster's short-term forecast becomes more informative ("predictive")

- $\Rightarrow$ But the forecaster's long-term forecast becomes less informative when:
  - Earnings are not too serially correlated
  - The cost of multitasking is high enough.

Forecast Informativeness

- **Definition:** The informativeness of the forecaster's forecast **at horizon h is**:

Prior uncertainty (earnings dispersion)

Posterior Uncertainty

$$R_h^2 = \frac{\mathrm{Var}(\theta_h) - \mathrm{Var}\left(\theta_h \middle| f_h^*\right)}{\mathrm{Var}(\theta_h)}$$

Normalization by prior uncertainty

Roadmap

Theory and predictions

Measuring analysts' forecasts informativeness

Long-run trend in analysts' forecasts informativeness and alternative data

Main test: New social media data

Take-away and implications

Equity Analysts' Forecasts

- We test these implications using sell side equity analysts' forecasts because analysts:

1. Are influential forecasters in securities markets
2. Regularly produce earnings' forecasts at various horizons
3. Increasingly rely on alternative data to produce their forecasts:

"*Nowadays, analysts sift through non-traditional information such as satellite imagery and credit card data or use machine learning and natural language processing to glean fresh insights […]*" (in "How investment analysts became data miners" Financial Times, November 28, 2019)

Measuring Analysts' Forecasts Informativeness

- Use U.S. sell-side analysts' earnings forecasts from I/B/E/S from 1983 to 2017
- For each analyst *i* on day *t*,
  - Retrieve latest forecast at horizon *h* for each firm *j* covered by the analyst
  - Regress in the **cross-section of covered firms** :

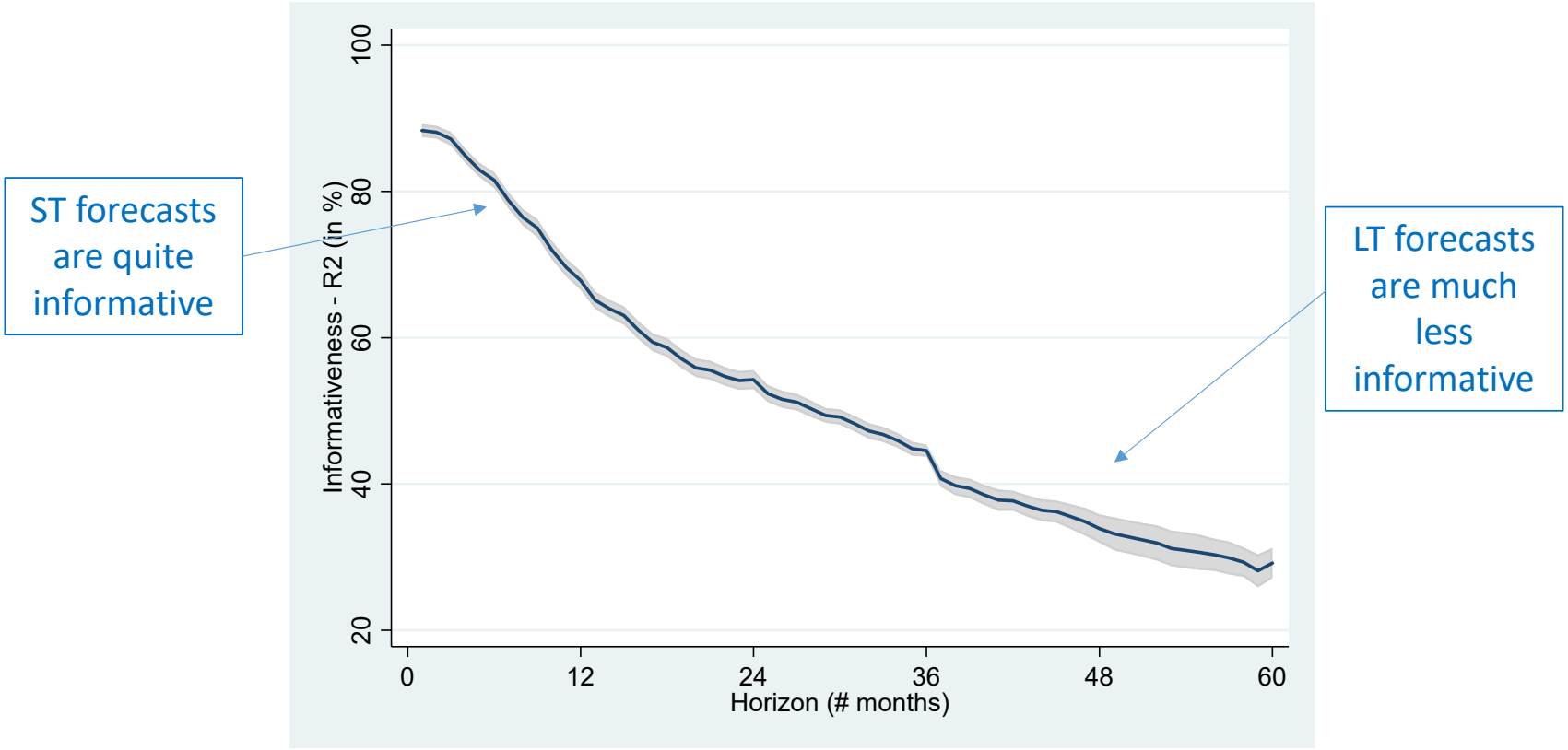
$$\frac{Earnings_{jh}}{Assets_j} = \alpha + \beta \frac{Forecasts_{jh}}{Assets_j} + \varepsilon_{jh}$$

  - … and estimate the R-squared $R^2_{i,t,h}$

- Higher $\mathrm{R}^2_{i,t,h}$ → More informative forecast at horizon h for analyst i.

Term-Structure of Forecasts informativeness

- 65,888,460 observations of analysts' forecasts informativeness (1983-2017).



Roadmap

Theory and predictions

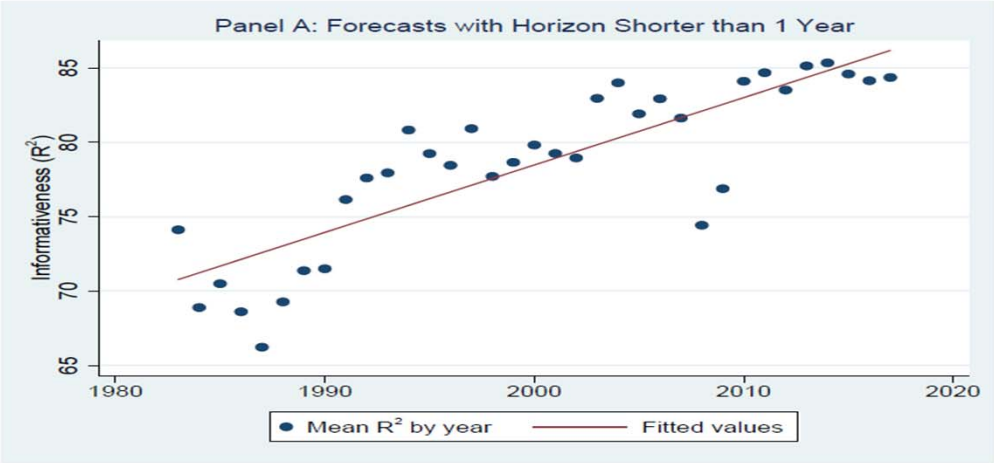
Measuring analysts' forecast informativeness

Long-run trend in analysts' forecasts informativeness and alternative data

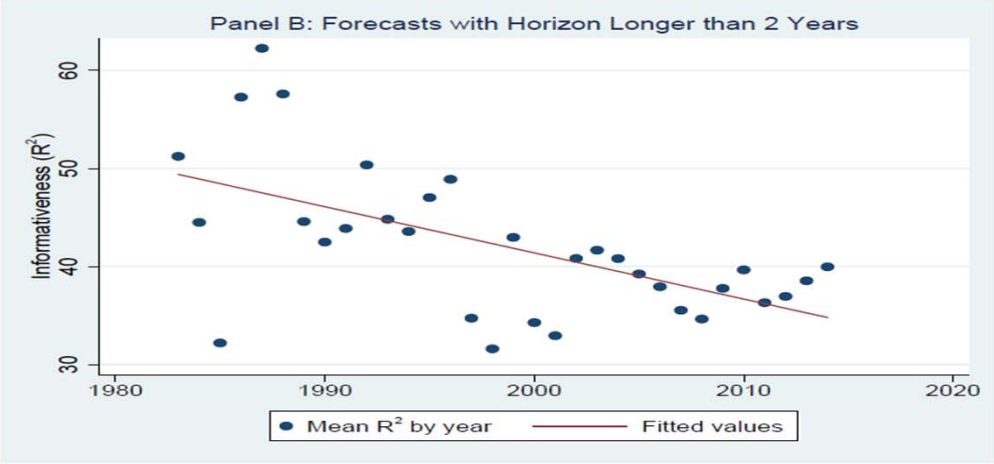
Main Test: New social media data

Take-away and Implications

Long Run Trend in Analysts' Forecasts Informativeness



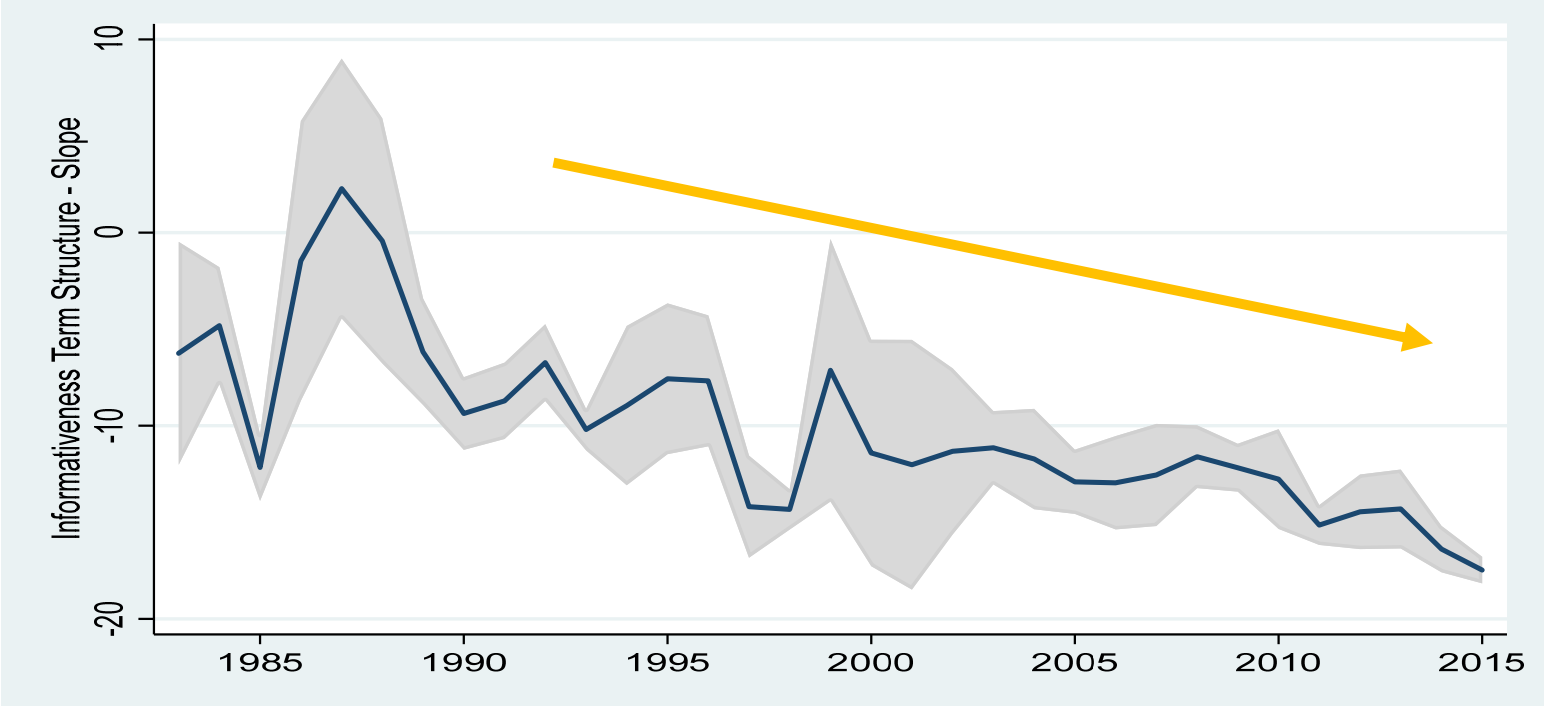
ST forecasts (1 year) become MORE informative



LT forecasts (>2 years) become LESS informative

Long Run Trend in Analysts' Forecasts Informativeness

- Estimate the trend in the "slope" of the term structure of forecast informativeness
  - **Every year**, estimate: $R^2_{i,t,h} = \mu_0 + \mu_1 \times h + \varepsilon$. Then plot μ_1



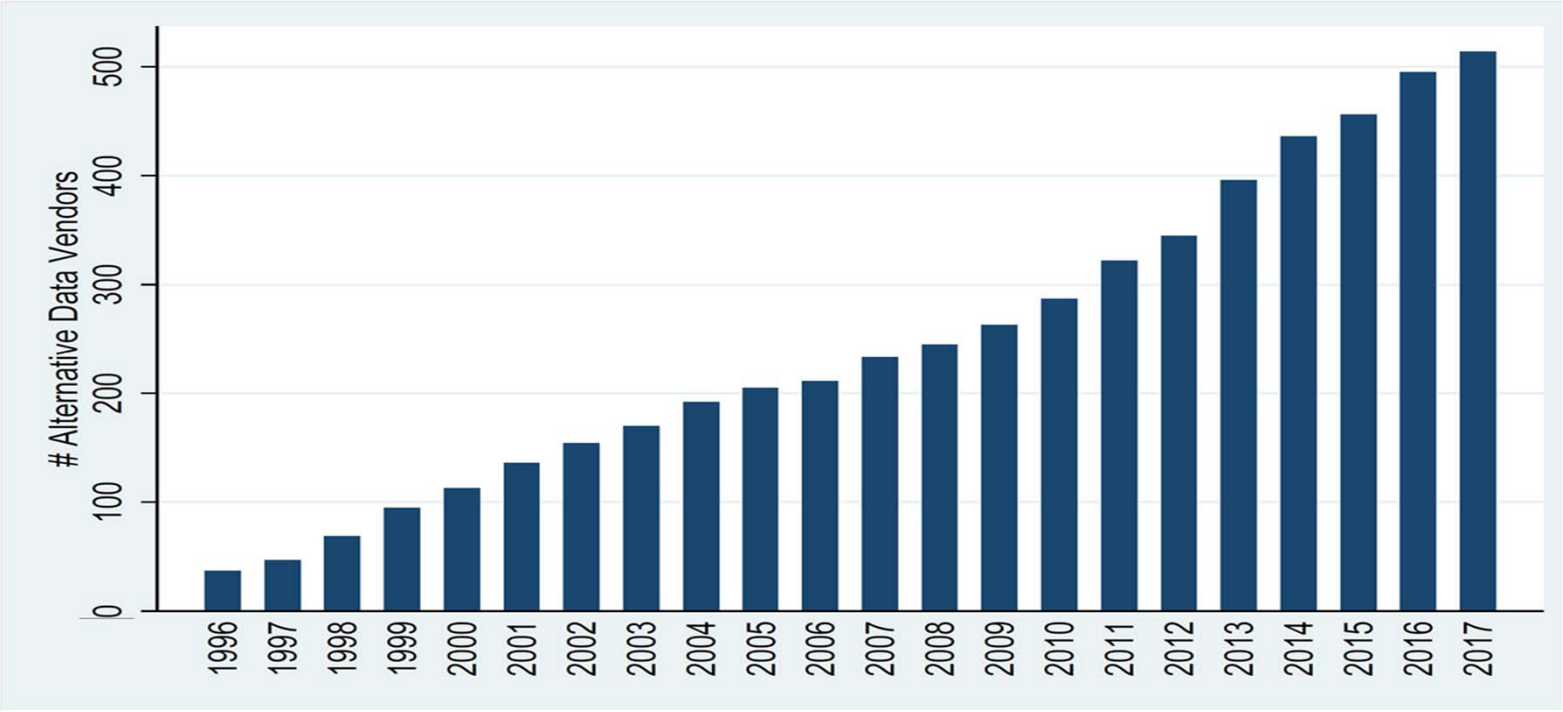
The slope becomes steeper over time: It shifts from about -6 (% per year) from 1983 to 1989 to -17 in recent years

Long Run Trend in Analysts' Forecasts Informativeness

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

The steepening of the term structure of analysts' forecasts informativeness holds when the slope is estimated by industry (FF17) or by analysts, and after controlling for industry or analysts fixed effects (hence the pattern is unlikely to be due to changes in sample composition).

Long Run Trend in the Supply of Alternative Data



- Are the two trends related? Our theory predicts that the steepening of the term structure of analysts' forecasts informativeness should be stronger for analysts using more alternative data. Is this the case?

The Alternative Data Channel

Dep. variable:	Estimated Year Trend in Slope by FF17			
WLS:	(1)	(2)	(3)	(4)
Alternative Data Usage 1		-15.9*		
		(-1.90)		
Alternative Data Usage 2			-1.2*	
			(-2.08)	
Alternative Data Usage 3				-0.6**
				(-2.18)
Constant	-4.2**	-0.5	1.5	2.3
	(-2.88)	(-0.16)	(0.48)	(0.61)
N	16	16	16	16
R^2	-	14.8%	28.1%	18.6%

- Cross-sectional regression of an estimate of the time trend in the slope of the term structure of analysts' forecasts informativeness by industry (FF17) on three measures of alternative data usage by analysts covering stocks in each industry over the 2000-2017 period (usage is measured by cites in analysts' reports of at least one data vendor in the 2019 JP Morgan directory of alternative data vendors)

Roadmap

Theory and Predictions

Measuring Forecasts' Informativeness

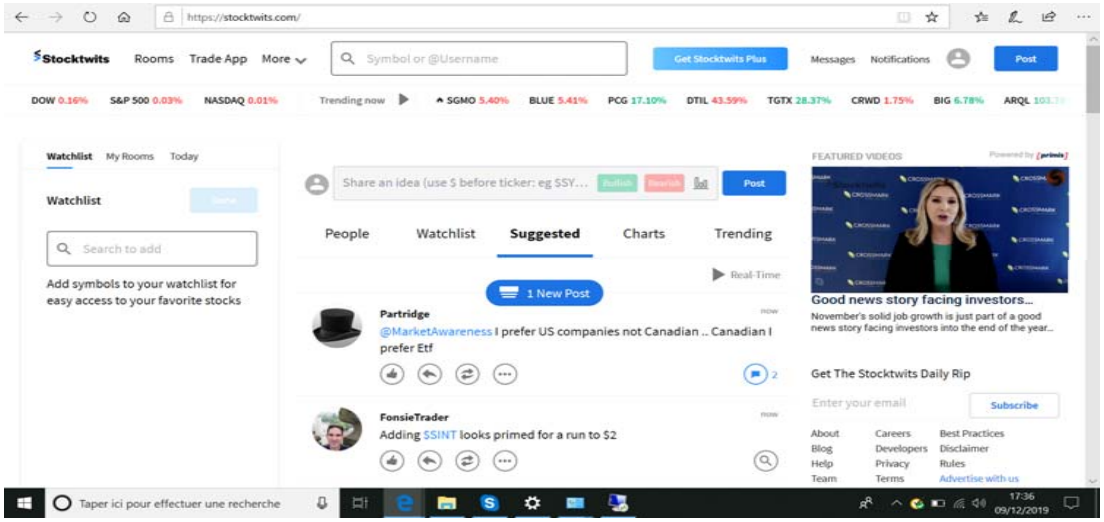
Long-run trend and alternative data usage

Main Test: New social media data

Take-away

StockTwits

- **StockTwits: The largest social network fully dedicated to U.S. financial markets (founded in 2009)**
  - Investors create posts with charts, links to articles, and opinions about stocks. Posts are linked to securities via a “cashtag”: $+TICKER



- **Our Data:** All messages posted on the platform from 2009 to 2017 (> 40 mio messages overall about 5,919 unique firms)

Why is StockTwits a good laboratory?

- **Equity analysts use StockTwits as a source of information:**
  - StockTwits has been integrated into all major aggregation platforms used by analysts (Bloomberg.com, Reuters.com, CNN Money, or Yahoo!)
  - We provide evidence that analysts are more likely to revise their forecasts about a stock following an increase in StockTwits messages about this stock (even after control for trading volume and in the absence of stock specific news in the past 30 days
  - Biographic information about analysts (I/B/E/S) and StockTwits' users' IDs: 35% of exact matches (based on 7,656 analysts during 2009-2017 period)

- **StockTwits contains short-term oriented information:**
  - Messages about a stock can be used to predict future earnings and sales but not beyond one year.

Empirical Design

- We estimate **for different horizons h**:

$$R^2_{i,t,h} = \alpha_{i,h} + \alpha_{t,h} + \boldsymbol{\lambda_h}\, \boldsymbol{Data\ Exposure}_{i,t-1} + Controls_{it-1} + \varepsilon_{t,j}$$

  - Analyst and day fixed effects (within-analyst variation over time)
  - Data Exposure=0 before 2009 (creation of the platform)
  - **λ_h measures the effect of exposure to social media data on the average analyst forecast informativeness for horizon h**
  - **Prediction (theory):**
    - $\lambda_h > 0$ for small h (Short Term)
    - $\lambda_h < 0$ for large h (Long Term)

Empirical Design

- Analyst i's exposure to StockTwits data on day t = Average Exposure to StockTwits of Stocks covered by Analyst i a date t.

- 2 measures of stock exposure to StockTwits (measured daily):

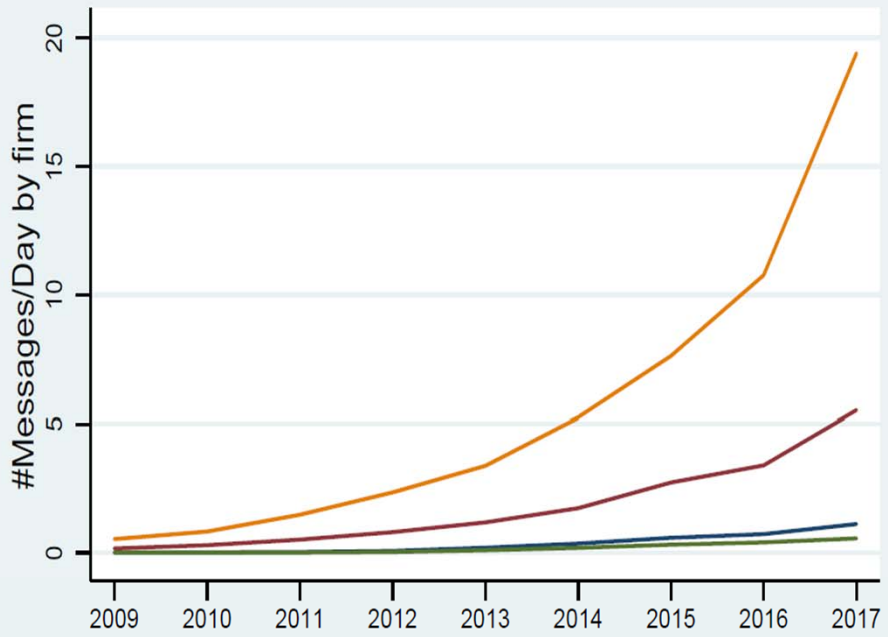
  1. **#Watchlist:**

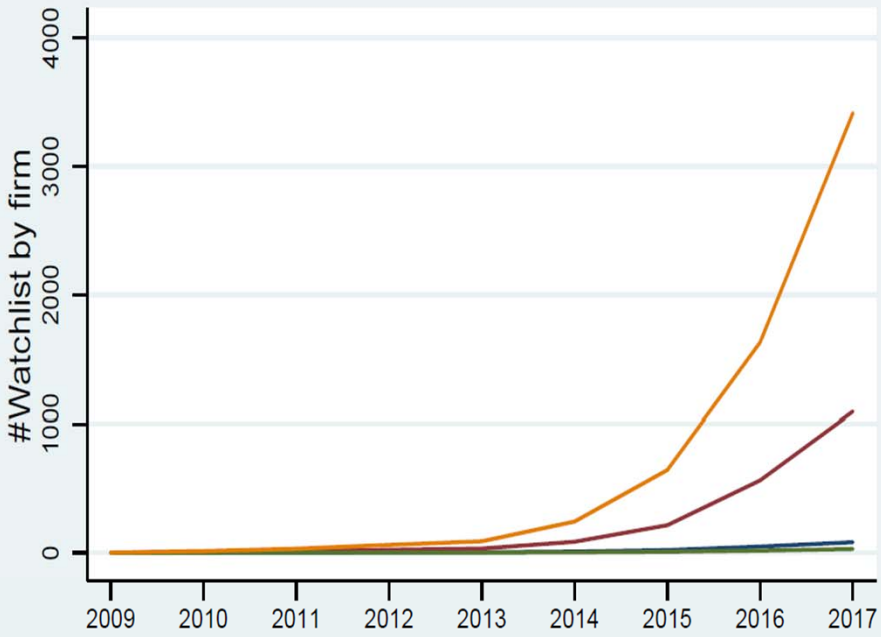
     - Number of users who declare following the stock when they register on StockTwits (orthogonal to news arrival from other sources)

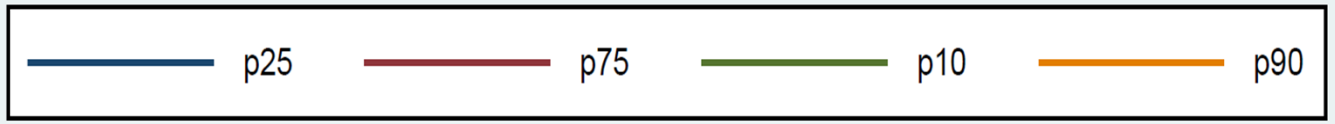
  2. **Hypothetical messages (averaged over 30 days)**

     - Total messages on StockTwits × average firm's "market share" of all messages
     - This share is persistent → **variation comes from variations in daily number of aggregate messages (**orthogonal to news arrival from other sources)

Heterogeneity in Analysts' Exposure







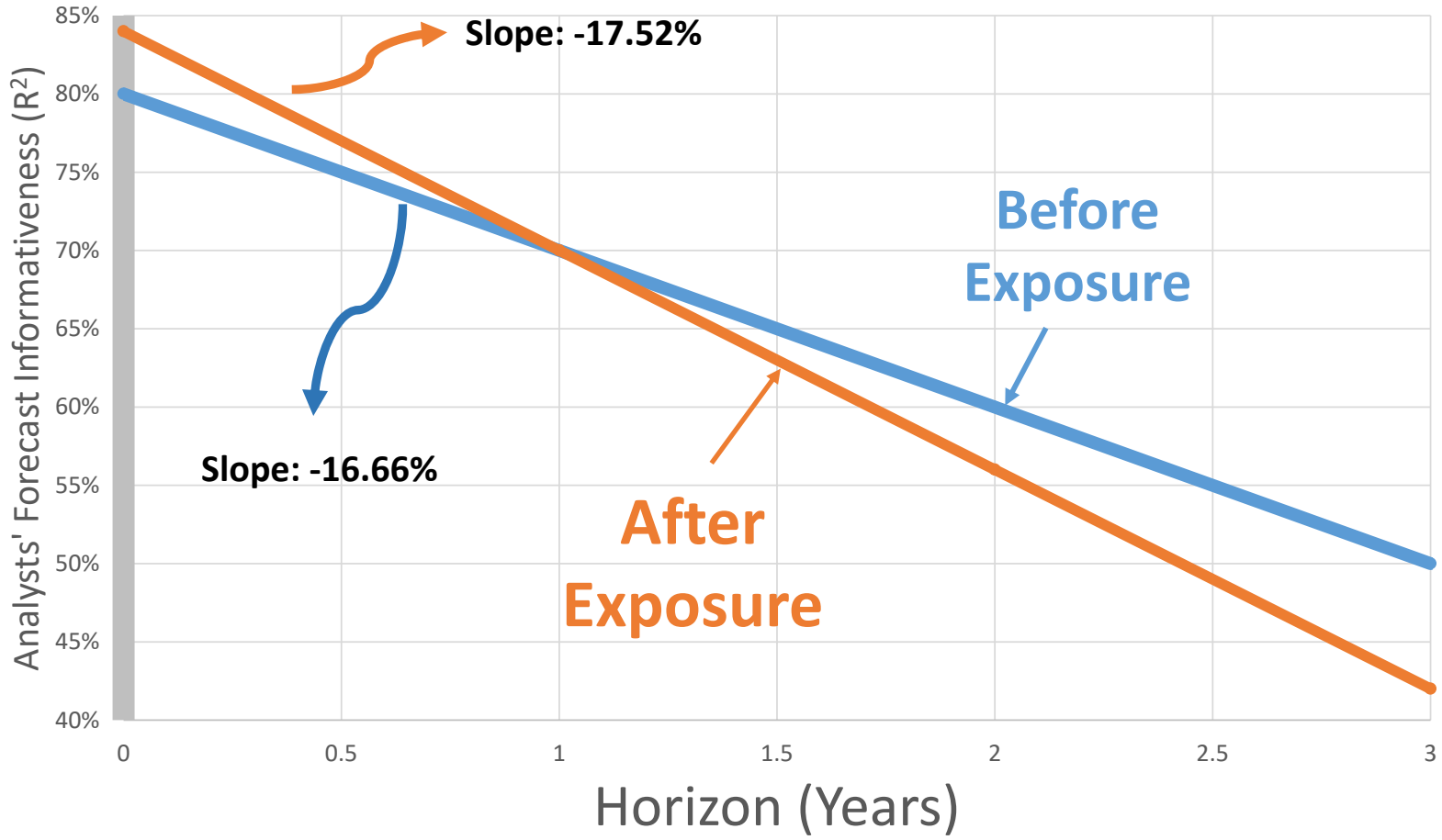
Main Finding

Dep. variable:	Short Term		Forecasts informativeness (R^2)		Long Term			
Sample:	$0 < h \leq 1$		$1 < h \leq 2$		$2 < h \leq 3$		$h > 3$	
OLS:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Proxy for Data Exposure = #Watchlist								
Data Exposure	0.54***	0.56***	0.4	0.23	-0.66***	-0.93***	-1.51***	-1.58***
	(3.89)	(4.13)	(1.06)	(0.63)	(-3.24)	(-4.78)	(-3.49)	(-3.27)
Analysts FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	14,055,963	13,026,178	11,489,986	10,596,175	3,916,280	3,634,874	1,496,954	1,434,373

ST forecasts become MORE informative

LT forecasts become LESS informative

The "StockTwits Effect"



Additional Findings

- **Other Predictions:** Deterioration in the informativeness of LT forecasts for an analyst **should be** more pronounced for:

    1. Analysts with a larger cost of multitasking is larger (measured by the number of stocks followed by an analyst)

    2. Analysts who follow stocks whose earnings are less correlated over time (ST info is less useful to predict LT earnings)

- We find evidence supporting both predictions.

Main Take-Away

- **Does alternative (short-term oriented) data improve securities analysts forecasts?**
  - –Yes at short horizon
  - – No at long horizon
- **More work is needed: Other forecasters, more direct measures of efforts in producing information (e.g., analysts questions during earnings calls etc.), other shocks to ST oriented data availability.**

Implications

- **Price informativeness:**
  - Stock price = sum of discounted forecasts of future cash flows $\Rightarrow$ Prices are informative about fundamentals.
  - If the informativeness of forecasts changes, the informativeness of stock prices ("distance" to fundamental") must change as well.
  - In which direction? If the informativeness of ST term forecasts increases but that of LT term forecasts decreases, the net effect is ambiguous.
  - May explain conflicting findings regarding long run trends in price informativeness (Bai et al. (2015), Farboodi et al. (2021)).

Implications

- **Real effects (corporate investment):**
  - If long-term forecasts become less informative, the value of long-term investment projects (projects whose cash-flows materialize in the long run) is reflected less quickly in stock prices.
  - $\Rightarrow$ Less incentive for managers to undertake long-term investment projects if managers' compensation depend on short-run stock prices ("short-termism").
  - $\Rightarrow$ Lower investment in long-term projects when the informativeness of long-term forecasts drop.
  - We provide evidence supporting this prediction in Dessaint, Frésard and Foucault (2022, work in progress).