Does Alternative Data Improve Financial Forecasting?
The Horizon Effect

Olivier DESSAINT   Thierry FOUCAULT   Laurent FRESARD

INSEAD   HEC Paris   USI Lugano and SFI

2021

Motivation

➢ Unprecedent increase in the volume and variety of data…

- E.g., Credit Card Transactions, Social Media Data, Satellite Images, App. Usage, Web Traffic Data

➢ … that may contain relevant information for financial forecasting

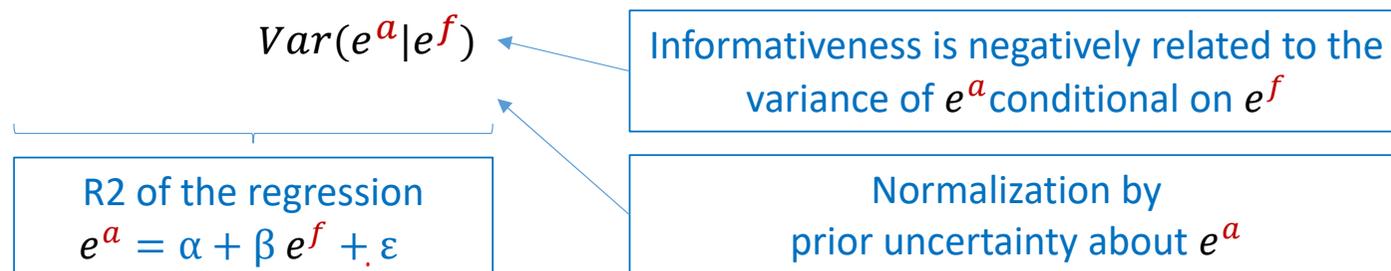➢ Research Question: **Have market participants become better at forecasting short and long-term cash flows? Why not?**

Our measure of forecasts quality by horizon h

➢ Develop a measure of forecasts informativeness using sell-side analysts earnings forecasts e^f (observable)

- e^f is informative if it reduces uncertainty about future actual earnings e^a

$$Var(e^a|e^f)$$

Informativeness is negatively related to the variance of e^a conditional on e^f

R2 of the regression
$$e^a = \alpha + \beta\, e^f + \varepsilon$$

Normalization by prior uncertainty about e^a

- Compute R^2 by regressing in the **cross-section of covered stocks j:**

$$\frac{e_j^a}{Assets_j} = \alpha + \frac{e_j^f}{Assets_j}\beta + \varepsilon_j$$

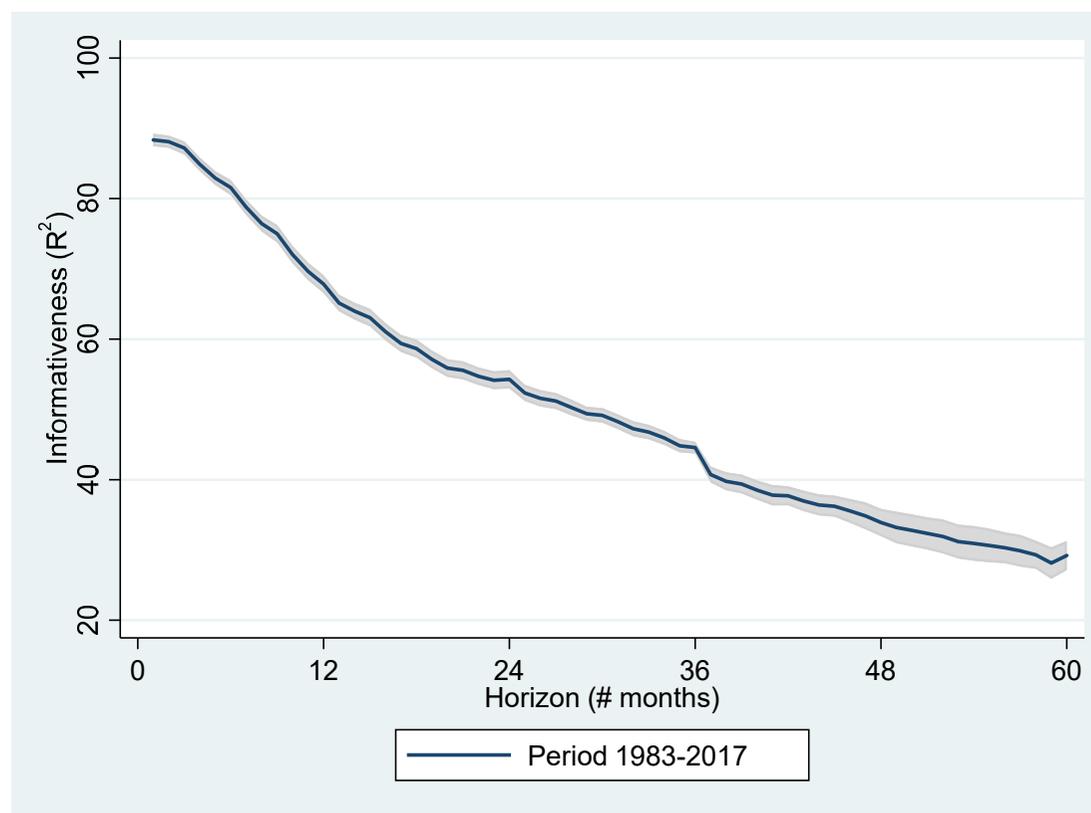- Obtain 65,888,460 obs. of R^2 by **analyst-day-horizon** (from **1983 to 2017**)

Term structure of forecasts informativeness

➢ Avg. R^2 ("informativeness") of analysts' earnings forecasts **by horizon**

High R^2 means forecasts explain a large fraction of variation in actual earnings
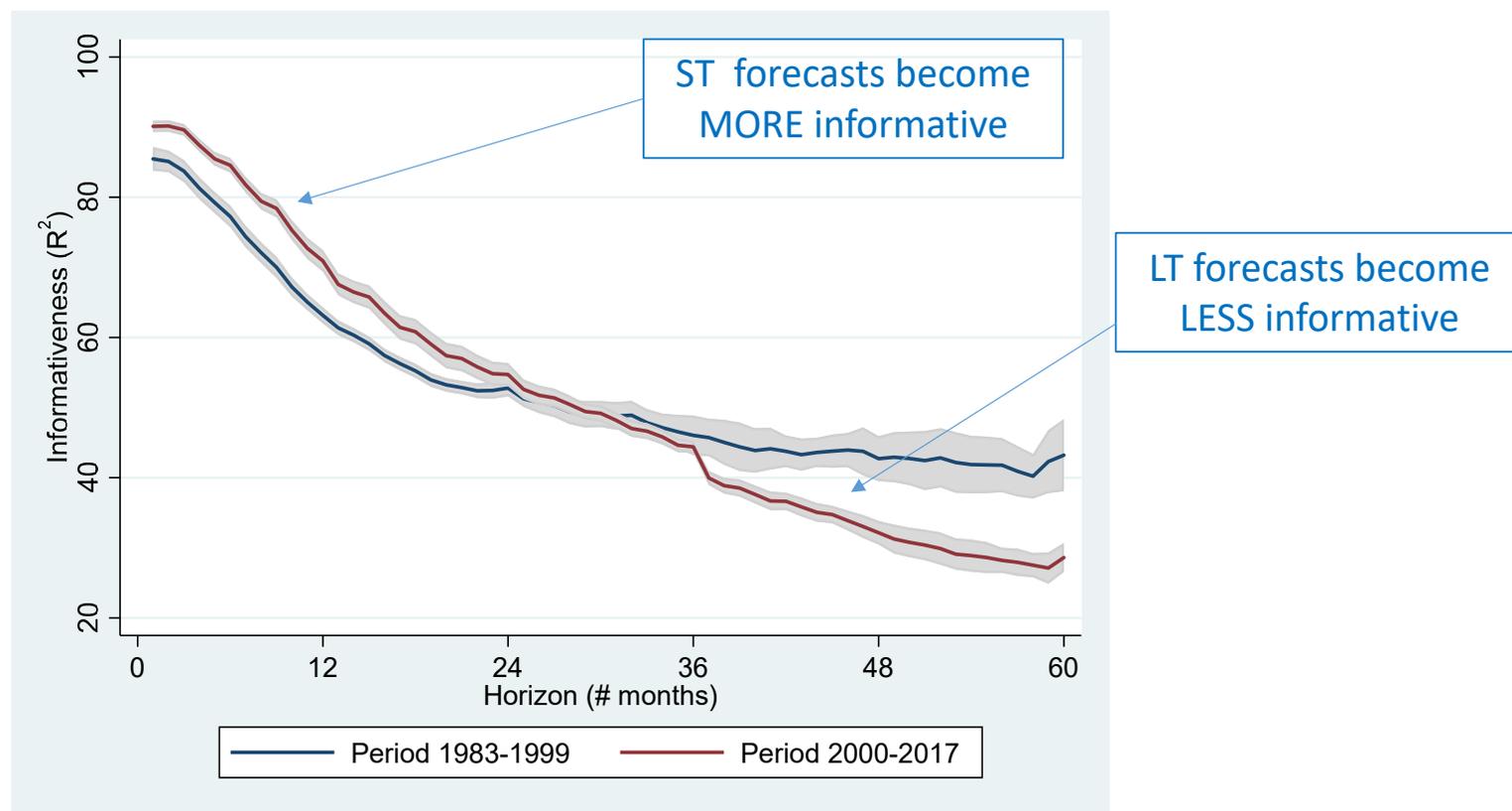
Average R2 decreases with horizon

Term structure of forecasts informativeness

➢ Avg. R^2 ("informativeness") of analysts' earnings forecasts **by horizon**

Term-structure has become steeper over time
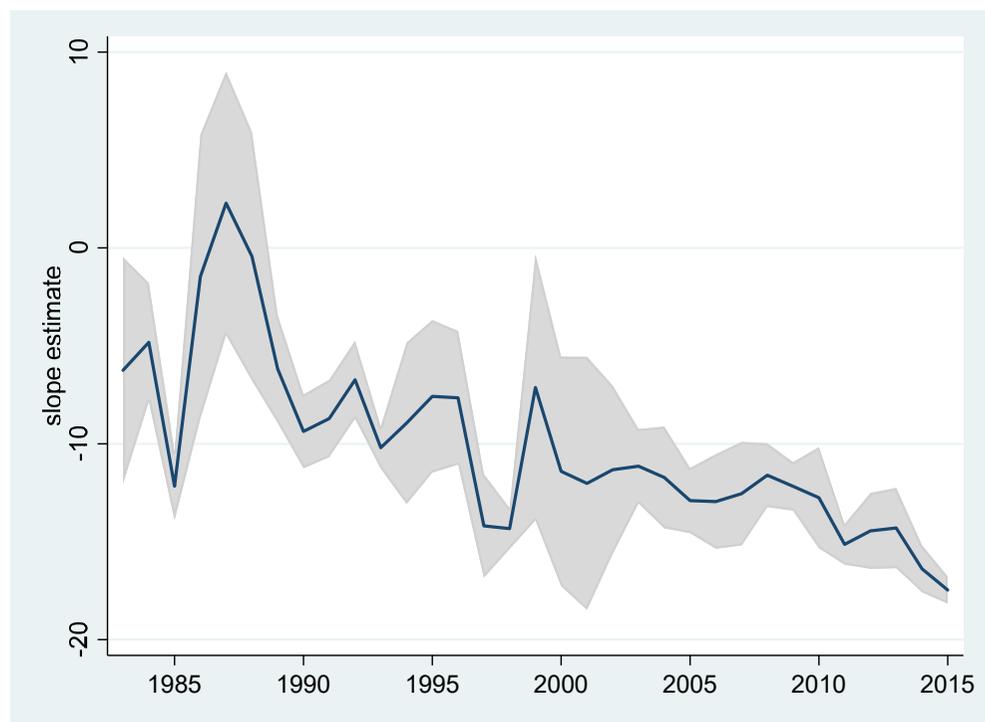
Term structure variation over time

- Estimate the <mark>slope</mark> of the term stucture by regressing R^2 on horizon $\boldsymbol{h}$

$$R^2 = \alpha + \beta \times \boldsymbol{h} + \varepsilon$$

- Do that every year and plot the slope (β) evolution

The slope becomes steeper over time
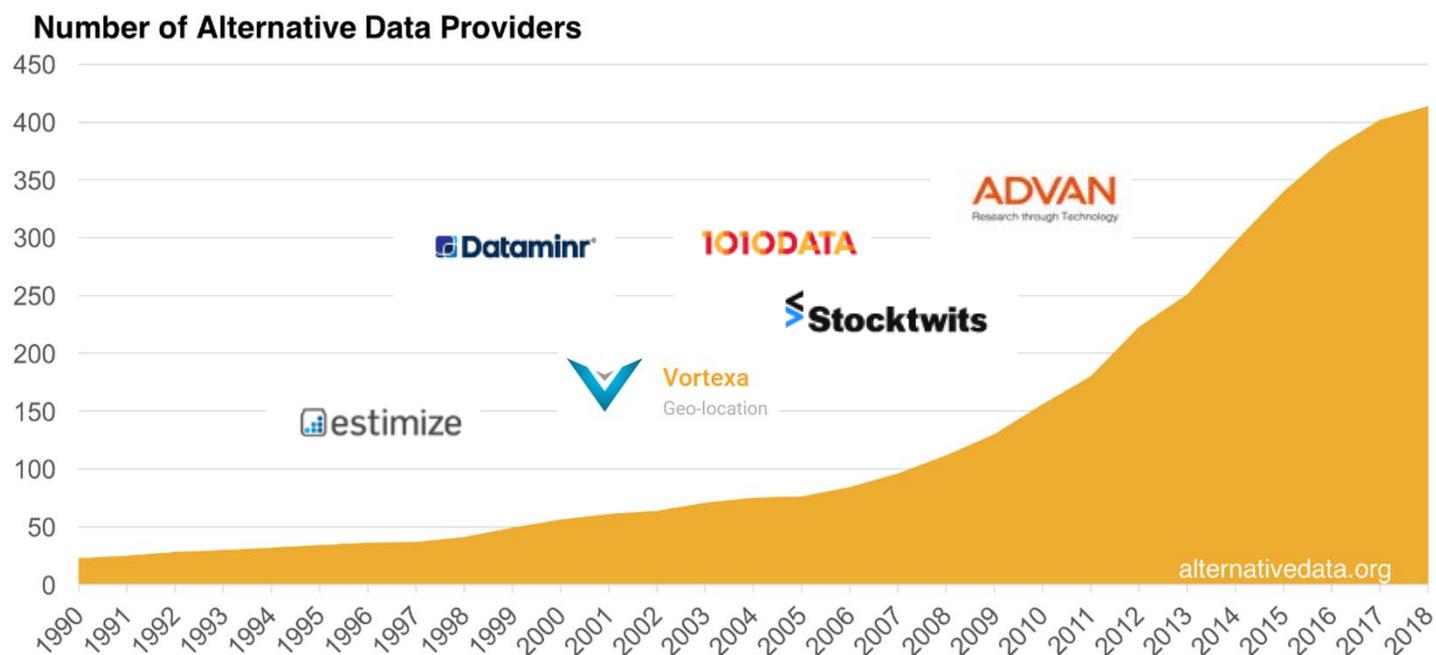
With an **acceleration after 2005**

Robust to formal testing with specifications including analyst and industry FE

Explanation?

➢ Shift in the term structure <u>coincides</u> with rise in alternative data

Many explanations for the previous trend are possible

We investigate only one

Number of Alternative Data Providers

➢ We now investigate the possibility that the two trends are related

Theory

Our theory: shift in term structure and alternative data are related

Investigate the implications those two arguments in a model

1) Alternative data are <u>relevant for short-term, mostly</u>

   - Alternative data have reduced the cost of producing information about ST relatively more than about LT

     - Why? Parking images or Web traffic data are useful to predict next quarter's earnings BUT not so useful to predict firms' long-term strategy and innovation

2) Forecasting ST and LT are <u>different tasks</u>. Switching from one task to the other is costly (<u>multitasking</u> problem)

   - ST (LT) cash flows depend more on assets in place (growth options)
   - Doing more of one task increases the marginal cost of doing the other

INSEAD The Business School for the World®

Model for production of information about ST and LT cash flows

➢ Key ingredients

- An analyst receives two noisy signals and precision depends on efforts

  ▪ 1 signal s_{st} about ST cash flow

  ▪ 1 signal s_{lt} about **_innovation_** in LT cash flow

  ▪ Cash-flow auto-correlation is known

  ▪ Utility (salary, promotion, …) decreases with average (squared) forecast error

- Total cost C of efforts z_{st} and z_{lt}

Agent minimizes expected forecast error net of information acquisition cost

$$C(z_{st}, z_{lt}) = az_{st}^2 + bz_{lt}^2 + c(z_{st}z_{lt}) \qquad \text{with a, b, and c > 0}$$

What happens when a ↓ (i.e. more alternative data)?

- For a given level of efforts:

  ✓ a = marginal cost for signal about ST

  ✓ b = marginal cost for signal about innovation in LT

  ✓ c = switching cost

The impact of alternative data in our model (Intuition)

➢ Predicted optimal efforts allocation with **alternative data (i.e. $a \downarrow$)**

1. Informativeness of short-term forecasts $\uparrow$

   BUT … doing more of one task increases the cost of doing the other (*c>0*)

2. Informativeness of long-term forecasts $\downarrow$

   IF …

   ✓ Switching cost *c* is high enough
   ✓ Cash flows auto-correlation is low enough

➢ Main prediction: **Alternative data can make short-term forecasts more informative and long-term forecasts less informative**

Empirical Design

➢ Our test: Use the introduction and expansion of StockTwits as a source of new info. about ST (i.e. $a \downarrow$)

- What is StockTwits?

  ▪ StockTwits is the largest social network fully dedicated to US financial markets. Discussion platform like « Twitter » but for traders only.

- Why is that a relevant laboratory?

  1. Social media data are used to extract signals (Chi et al. (2021)

     ✓ StockTwits datafeed integrated to Bloomberg and Reuters.

     ✓ Up to 30% of analysts could have an account (based on name matching)

  2. Opinions are are relevant for predicting short-term cash flows

     ✓ 80% of messages are by "day-traders" and "swing-traders"

     ✓ "Buy" or "Sell" reco. by users predict short BUT NOT long-term cash flows

  3. Created in 2008. Coverage expanded progressively with different level of intensity across US stocks after 2009

Methodology

➤ Use a « diff-in-diff like » estimation at the analyst level

$$R^2_{i,t,h} = \alpha_t + \alpha_i + \gamma \ Data \ Exposure_{i,t} + \varepsilon_{i,t}$$

"Watchlist" is a **sticky** list of stocks defined by the user at the time of first registration

➤ *Data Exposure* = Average # of users that have the stocks the analyst covers in their **watchlist**

- ▪ Equal to 0 before 2009 (test starts 5 years before introduction)
- ▪ Normalized by standard-deviation

"Watchlist" is **NOT** correlated with news arrival from **other** sources of info.

➤ γ measures change in R^2 after introduction for analysts highly exposed relative to change in R^2 for analysts less exposed or with staggered exposure

Main results

ST forecasts become MORE informative

LT forecasts become LESS informative

Panel A - Proxy for Social Media Exposure : # Watchlist

	Dependent variable: Forecast informativeness (R2 measure in percentage points)							
OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	h <=1 Yr		1 Yr < h <=2 Yrs		2 Yrs < h <=3 Yrs		h>=3 Yrs	
Data Exposure	0.54***	0.53***	0.4	0.18	-0.65***	-1.00***	-1.51***	-1.55***
	(3.90)	(4.03)	(1.07)	(0.47)	(-3.20)	(-4.78)	(-3.49)	(-3.20)
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	14,026,800	13,006,543	11,502,199	10,612,608	3,929,446	3,648,151	1,500,165	1,438,756

Ancillary results

➤ We also find

✓ Stronger effect when **switching is more costly** (Parameter c is higher)

✓ Lower effect when **auto-correlation is high**

Consistent with our hypothesis of a change in the cost function of information acquisition, driven by alternative data

✓ Stronger effect when cost of ST info. decreases more (Parameter a is lower)

➤ Many more tests, results, discussions are in the paper ! …

Conclusion

➢ Shift in forecasts informativeness over the past decade

- • Better at short-horizons and worse at long-horizons

➢ We relate this trend to recent increase in alternative data

- • Alternative data provide a "comparative advantage" for making ST forecasts

- • Doing more of one task (ST) increases the cost of doing the other (LT)

- • ST forecasts can become more informative at the expense of LT forecasts

➢ Empirical evidence from the introduction of a major social media provides support to this hypothesis