THE USE AND USEFULNESS OF BIG DATA IN FINANCE

Feng Chi, Byoung-Hyoun Hwang, and Yaping Zheng*

This Draft: August 2022

We study to what degree analysts draw from "big data." To measure analysts' reliance on big data, we record how frequently they reference the use of big data in their written reports. We find that analysts frequently draw from big data, particularly when receiving timely signals regarding a company's performance is important, when traditional data are ambiguous, and when analysts lack access to granular data. Analysts' reliance on big data is accompanied by notable improvements in forecast accuracy and heightened stock-market reactions to their forecast revisions.

JEL Classification: G12, G14, G17, G24.

Keywords: Big Data, Alternative Data, Sell-Side Analysts, Performance.

---

* Chi is affiliated with the Cornell SC Johnson College of Business, Cornell University. Hwang is affiliated with Nanyang Business School, NTU, Singapore. Zheng is affiliated with the Desautels Faculty of Management, McGill University. E-mail: fc354@cornell.edu, bh.hwang@ntu.edu.sg, and yaping.zheng@mail.mcgill.ca. The authors would like to thank Andy Call, Clifton Green, Jillian Grennan (discussant), Allen Huang (discussant), Jiekun Huang, Russell Jame, Andrew Karolyi, Ningzhong Li, Ross Lu, Stan Markov (discussant), Joe Pacelli, Marcus Painter (discussant), Hongping Tan, Ye Wang (discussant), Scott Yonker, Chi Zhang (discussant), Frank Zhang, Jingjing Zhang and seminar participants at Korea University, 2021 EFA Annual Meeting, 2021 MIT Asia Conference in Accounting, 2021 China International Conference in Finance, 2021 CAPANA Conference, 3rd Future of Financial Information Conference, and 2021 Conference on Financial Innovation for many valuable suggestions and comments.

1. Introduction

With the advent of modern information technologies and recent advances in data analytics, we can increasingly track individuals' and businesses' activities through the digital footprints they leave behind. A large body of research suggests that this kind of new data, also known as "big data" or "alternative data," can help a variety of users make better decisions.[1] Examples range from companies using alternative data to better predict product demand and manage their inventories (Mayer-Schönberger and Cukier, 2013) to hospitals drawing from alternative data to better anticipate the number of patients (Ambert et al., 2016).

A growing literature suggests that alternative data can also help investors better evaluate companies. For instance, Huang (2018) shows that customer product reviews on Amazon.com provide insights into a company's future performance. Green, Huang, Wen, and Zhou (2019) and Huang, Li, and Markov (2020) find that how employees rate their employers on Glassdoor.com positively predicts firm performance.

Corresponding accounts in the financial press suggest that investors increasingly draw from alternative data (Dannemiller and Kataria, 2017; Ram and Wigglesworth, 2017; Watts, 2019; Wigglesworth 2020). By 2022, investors are estimated to spend more than three billion a year on the acquisition of alternative data (PRNewswire, 2022).

While alternative data may provide valuable insights, the high cost and the resulting unequal access to the data can also have large negative consequences. In the financial domain, it is commonly assumed that the key beneficiaries of alternative data are hedge funds and that "mainstream Wall Street" and ordinary investors do not draw from alternative data (Dannemiller and Kataria, 2017; Watts, 2019). This raises the prospect that the arrival of alternative data has un-leveled the playing field among investors, thereby leading to greater information asymmetry and lower financial market quality (Glosten and Milgrom, 1985; Easley and O'Hara, 1987; Diamond and Verrecchia, 1991).

[1] In finance, the terms "big data" and "alternative data" are used interchangeably (e.g., Savi, Shen, Betts, and MacCartney, 2015; Whyte, 2017). Others use the term "big data" more broadly to describe a combination of data/method and a new set of principles. For instance, Mayer-Schönberger and Cukier (2013) describe "big data" as "*the ability of society to harness information in novel ways to produce useful insights or goods and services of significant value*." In this study, we almost exclusively use the term "alternative data."

Our paper examines this assertion and provides direct evidence to what degree mainstream Wall Street draws from alternative data. We also consider performance implications. More broadly, our paper provides new systematic evidence regarding market participants' use of alternative data, what types of alternative data they use most frequently, and in what situations.

The group of market participants we consider in this study are sell-side analysts. Analysts routinely publish written reports in which they describe their opinions as well as the data and analyses they used to arrive at their views. If the consideration of alternative data meaningfully altered an analyst's belief, it appears plausible that the analyst would discuss this in the corresponding report. We thus posit that by parsing analysts' written reports and counting how often they explicitly reference the use of alternative data, we can gauge to what degree analysts draw from alternative data.

To provide some details on our parsing approach, we start with a comprehensive list of in-house data science teams and "external" alternative-data vendors. We search for the names of these teams and vendors in analysts' written reports. We then conduct an iterative keyword search following prior literature (Hoberg and Moon, 2017, 2019). Among the reports that contain the name of a team or vendor, we extract a list of keywords that analysts use to describe the alternative data. We then use these keywords to search for additional reports that draw from alternative data and expand our list of keywords. Through these iterations, we arrive at our final set of keywords, which we use to identify reports that explicitly reference the use of alternative data. At the end of our process, we manually read the relevant passages of each captured report to verify that the report indeed draws from alternative data.

Given the labor intensity of our identification process, we restrict our analysis to constituents of the Dow Jones Industrial Average index (DJI). Our final sample comprises 64,036 written reports compiled by 1,002 distinct analysts working for 55 brokers from June 2009 through May 2019.

Our analysis reveals that by 2009/2010, 11% of the analysts in our sample explicitly reference the use of alternative data in at least one of their reports. By 2018/2019, the corresponding fraction is 28%. Our analysis differentiates between eight alternative data categories: app usage, sentiment, employee, geospatial, point of sale, satellite image, web traffic, and others. We find explicit references from all eight categories within the first year of our sample period. Overall, our evidence shows that analysts frequently draw from alternative data, at least for the largest and economically most meaningful firms. Our evidence also shows that they have been doing so for a long time.

Examining analysts has the appealing feature that we can directly observe their beliefs through the earnings forecasts they issue. More importantly, we have an objective benchmark, in the form of actual reported annual earnings, against which the earnings forecasts can be compared. Combined, these features allow us to gauge whether the reliance on alternative data has helped analysts form more accurate beliefs.

We find that when analysts draw from alternative data, they issue substantially more accurate earnings forecasts. Our estimates suggest that the performance improvement that accompanies the consideration of alternative data is equivalent to having covered the corresponding firm for 3.7 additional years.

We evaluate three possible sources of alternative data's seeming usefulness. Compared with traditional data, alternative data can provide more timely signals regarding a company's performance. In addition, since alternative data are compiled by third-party vendors, firms are removed from the data-generating process, thereby mitigating concerns of misrepresentation.[2] Finally, alternative data provide signals at a highly disaggregated level, such as at the product or branch level. The granularity of alternative data may give investors a more nuanced picture of a company's performance.

Our results suggest that all three sources are important. We find that analysts more frequently draw from alternative data when it becomes more relevant to receive instantaneous signals regarding a company's performance, specifically, when a company provides few official updates and when a company's earnings are highly volatile and difficult to forecast. We also find that analysts more frequently draw from alternative data when traditional data are ambiguous, specifically when a firm has had to restate its financial accounts and when a firm uses high levels of discretionary accruals. Finally, we find that analysts more frequently

[2] Relatedly, Mukherjee, Panayotov, and Shon (2021) find evidence that commercial satellite images provide more timely signals about the macroeconomy than "traditional" government announcements. The authors suggest their "*results point to a future in which the resolution of macro uncertainty is smoother and governments have less control over macro information*" (abstract).

draw from alternative data when they lack preferential access to management and thus cannot easily obtain performance signals at a granular level.

Our final analysis considers stock-market reactions to the use of alternative data. We find that stock-price changes corresponding to changes in earnings forecasts, changes in target prices, and changes in recommendation levels are roughly two to three times larger when an analyst explicitly references the use of alternative data compared to when an analyst does not. These differential stock market reactions do not revert. Our results suggest that alternative data insights become priced.

Our paper is part of a growing literature that examines the consequences of alternative data for financial markets. The current literature mostly shows that alternative data provide valuable insights into companies' performances (Bartov, Faurel, and Mohanram, 2018; Huang, 2018; Green, Huang, Wen, and Zhou, 2019; Huang, Li, and Markov, 2020; Agarwal, Qian, and Zou, 2021; Dichev and Qian, 2022; Gupta, Leung, and Roscovan, 2022; Sheng, 2022). To what degree investors actually use and benefit from alternative data and the corresponding ramifications for investors and financial markets is less clear. The fact that much of the literature finds that alternative data predict earnings above and beyond those forecasted by sell-side analysts provides indirect evidence that while hedge funds may use and benefit from alternative data, mainstream Wall Street does not.

Our paper shows that analysts do incorporate alternative data and that the stock market does pay greater attention to analysts who draw from alternative data. Our evidence thus suggests that the emergence of alternative data has helped broad sections of the investor population make better investment decisions.[3]

Our paper also provides new systematic evidence regarding the types of alternative data that an important group of financial market participants uses and the factors that increase the likelihood of alternative data adoption. These descriptive statistics should serve as a useful reference for future work on

[3] How can we reconcile our evidence that analysts incorporate alternative data with prior literature's finding that alternative data predict earnings surprises? Prior literature considers a broad cross-section of firms, including smaller firms, for which the trading capacity and incentives to uncover possible mispricing are low. Our study focuses on the largest firms. The predictability of earnings surprises noted by prior literature could be driven by the smaller firms and may thus say more about incentives than about analyst sophistication.

this topic. These descriptions also shed light on the underlying forces that make alternative data particularly helpful in the investment domain.

2. To What Degree Do Analysts Draw from Alternative Data?

We begin the main body of our paper with a definition of alternative data. We then introduce our method of capturing analysts' reliance on alternative data and present the corresponding descriptive evidence.

2.1 Alternative Data and Historical Perspective

Alternative data trace the footprints individuals and firms leave behind through their day-to-day activities. These footprints are commonly referred to as "exhaust data." The use of exhaust data is not new to the financial sector. In the past, investment firms dispatched their junior analysts to retail stores to sample the foot traffic; other firms directed their analysts to manufacturing plants to count the number of trucks moving in and out (McMahon and Chu, 2012; Wigglesworth, 2016).

What distinguishes alternative data from the previous exhaust data is that with the advent of modern information technologies and the rise in computing power, we can now source exhaust data instantaneously, comprehensively, and from a variety of sources. That is, rather than manually count the foot traffic for select branches over a few days, we can now comprehensively track how many consumers visit a merchant's website.

There are broadly eight alternative data categories: (1) app-usage data, which track the number of active mobile app users, and the amount of time they spend on the apps; (2) sentiment data, which include product ratings posted on the Internet and social-media feeds regarding a company's products and services; (3) employee data, which include online job postings, employee opinions, and manager statements; (4) geospatial data, which contain information about the locations in which a company operates branches; (5) point-of-sale data, which include merchant-level transaction data, product-level purchase data, and pricing data; (6) satellite-image data, which include satellite images of parking lots, manufacturing plants, and construction sites; (7) web-traffic data, which track what terms users search for in the Internet and how frequently and for how long users visit a merchant's website; and (8) other, which include data that do not fit cleanly into any of the other seven categories (e.g., bills of lading detailing the type, quantity and destination of goods in a shipping container).

2.2 Measuring the Reliance on Alternative Data

To measure analysts' reliance on alternative data, we use Investext to download all sell-side analyst reports for DJI constituents from June 1, 2009, through May 31, 2019. DJI constituents represent 30 large publicly traded firms. Because the DJI constituent list varies over time, our final sample comprises 35 firms.[4] For each report, we extract the ticker symbol, company name, report date, the analyst names, the name of the broker, the report title, and the full text. The average report in our sample contains 2,152 words, which is the equivalent of roughly five pages.

We merge our Investext data with annual earnings forecast data from the IBES database. We merge these two datasets based on ticker symbols, company names, broker names, analyst names, and dates of forecast issuances.

Out of our initial sample of 70,353 Investext reports, we successfully match 65,009 Investext reports (65,009/70,353 = 92.4%). After merging with financial market data from CRSP and financial statement data from Compustat, our final sample comprises 64,036 reports and earnings forecasts issued by 1,002 analysts from 55 brokers covering 35 firms.

We then proceed as follows: We compile a list of in-house data science teams and external alternative-data vendors from the vendor lists in the J.P. Morgan 2019 Alternative Data Handbook and AlternativeData.org, a platform that connects users to alternative data providers.[5] To facilitate research on

[4] The mean and median market capitalization of firms in our sample (as of 2019) are $250 billion and $222 billion, respectively. To put these numbers in perspective, the 99th market capitalization percentile among firms in the CRSP/Compustat universe (also as of 2019) is $144 billion. DJI constituents are thus substantially larger than most firms in the CRSP/Compustat universe. Online Appendix Table A1 also compares the industry distribution of the firms in our sample with that of the firms in the CRSP/Compustat universe. Compared with the CRSP/Compustat universe, our sample overweighs the Consumer Staples sector and underweighs the Health Care sector.

[5] In-house data science teams specialize in collecting and analyzing large unstructured data, which analysts can use in their valuation efforts.

this topic, we report the list of 520 teams and vendors in Online Appendix Figure A1. We use the list of full and abbreviated names as our initial keywords and search for them in analysts' written reports.[6] For each report identified by these initial keywords, we read the passages surrounding the initial keywords to verify that the report indeed relies on alternative data.

Some analysts explicitly reference the use of alternative data without disclosing their source. To capture these reports, we follow prior literature (Hoberg and Moon, 2017; 2019) and conduct an iterative keyword search process. In particular, within our first set of analyst reports that adopt alternative data and reference an in-house data science team or external alternative data vendor, we extract a list of keywords that analysts use to describe the alternative data. We then use these new keywords to search for additional reports that draw from alternative data (but do not reference their source) and continue expanding our keywords list. Using this iterative process, we arrive at our final set of keywords, which we use to identify reports that explicitly reference the use of alternative data. We report our final set of keywords in Appendix 1. In our last step, we (again) read all reports flagged as using alternative data to verify that the analysts indeed rely on alternative data in their analyses.[7]

Our approach misses instances where analysts draw from alternative data but do not discuss their reliance in their reports. However, we are confident that all the reports that we mark as using alternative data indeed draw from alternative data. The fraction of times that analysts explicitly reference the use of alternative data is thus a *downward*-biased estimate of analysts' true reliance on alternative data.

To illustrate our process by example, one of the alternative-data vendors in our sample is "Remote Sensing Metrics," also referred to as "RS Metrics." We first search for reports containing the terms "Remote Sensing Metrics" or "RS Metrics." We find 47 reports that contain these two keywords. The figure below is an excerpt from one such report:[8]

[6] We convert all names and all text in the reports to lowercase characters.
[7] To researchers interested in further studying the use of alternative data, we would like to caution that analysts' use of the keywords presented in Appendix 1 is a necessary but not a sufficient condition. We found our final step of carefully re-reading all reports to eliminate false positives crucial in cleanly separating reports that rely on alternative data from those that do not.
[8] UBS; Neil Currie, Krista Zuber, and David Eads; Walmart Inc; August 12, 2010.

Wal-Mart Stores | 12 August 2010

UBS Proprietary National Parking Lot Fill Rate Analysis

We have conducted an analysis with Remote Sensing Metrics, LLC to track parking lot fill rates in order to predict overall US comp-sales performance at Walmart Stores using a sample of between 100 and 150 like-for-like satellite images each month for the past six months. Samples are representative of geographic region, store formats, day of week, and the time of period analysis. All satellite images are usually taken between 10:30am and 1pm to minimize shadows on the images. We believe a traditional grocery trip is less fixed to a certain time of day and thus the time-slot window for imagery results bears less risk than for other more discretionary shopping trips.

Reading the text surrounding the keyword "Remote Sensing Metrics," we identify two additional keywords related to alternative data: "parking lot fill rates" and "satellite image." We use these new keywords to search for more reports that rely on alternative data but do not reference "Remote Sensing Metrics" or "RS Metrics."

The figure below is an excerpt from one such report:[9]

- **April Results In +LSD Range:** The McDonald's U.S. sales result for April was +4.0%, which we believe included no meaningful menu price benefit. For April, we had predicted SSS of +3.0%, based widely on our proprietary parking lot fill rate analysis which had suggested lunch trends were positive and in the +3.0% range. We believe that featured core products, breakfast, and beverage platforms marginally contributed to overall U.S. comp trends in line with our expectations. Backing into our quarterly estimate, our +3.3% domestic 2Q11 projection suggests a +3.0% estimated comp for May and June.

In our final step, we read all reports that our procedure flags as using alternative data to verify that the analysts indeed draw from alternative data in their analyses. For instance, some firms in our sample provide satellite-related products or employ satellite imagery in their business processes (e.g., oil and gas exploration). We exclude such cases. The figure below is an example of a false positive:[10]

[9] Piper Jaffary; Nicole Miller Regan and Joshua C. Long; McDonald's Corporation; May 9, 2011.
[10] Jefferies Group. John DiFucci, Joseph Gallo, and Howard Ma; Microsoft Corporation; May 11, 2017.

> **Digital Transformation.** Much of the Analyst day was focused on capturing customer relevance which will translate to revenue growth. The company believes that its TAM has increased to $ 4.5 trillion today. In one example, Microsoft helped Land O'Lakes with a multiple year digital transformation which saw the company use Office 365, Surface, Windows 10, Azure, and Hololens. They were able to take decades of satellite imagery, load it into Azure, and then analyze the land, different weather metrics and type of seeds to better layout farms. This process drove yield increases from 130 bushels of corn per acre to now over 500 bushels.

Our primary variable, *I(Alternative Data*$_{i,f,t}$*)*, equals one if analyst *i*'s forecast for the annual earnings of firm *f* at time *t* is accompanied by a written report that explicitly references the use of alternative data and zero otherwise.

2.3 Descriptive Evidence Regarding Analysts' Reliance on Alternative Data

Our first test examines how much analysts draw from alternative data and, if so, in what manner. Panel A of Table 1 reports summary statistics for *I(Alternative Data)* across years. Since we have only partial data for 2009 and 2019, we combine the observations in 2009 with those in 2010 and the observations in 2019 with those in 2018. Panel A reveals that in 2009/2010, 6% of the analyst forecasts are supported by alternative data. By 2018/2019, 10% of the forecasts draw from alternative data.

The fraction of analysts drawing from alternative data for at least one of the firms they cover is naturally larger than the fraction of reports discussing alternative data use. In particular, we find that by 2009/2010, 11% of the analysts in our sample have drawn from alternative data. By 2018/2019, the corresponding fraction is 28%. As alluded to in Section 2.2, we believe this fraction understates analysts' true reliance on alternative data.

Panel B reports summary statistics for *I(Alternative Data)* across industry sectors. The use of alternative data is most widely referenced for firms in the Information Technology sector: the average *I(Alternative Data)* is 16%. It is also widely referenced for firms in Consumer Discretionary (10%), Consumer Staples (10%), Communication Services (9%), Health Care (8%), and Industrials (6%). It is infrequently referenced for firms in Energy (2%), Financials (1%), and Materials (1%).

In our study, we manually assign reports that draw from alternative data into the following eight categories: (1) app-usage data, (2) sentiment data, (3) employee data, (4) geospatial data, (5) point-of-sale data, (6) satellite-image data, (7) web-traffic data, and (8) other types of alternative data. Some reports are assigned to more than one category as analysts occasionally draw from multiple alternative-data categories. Appendix 1 details how we allocate the reports to the above eight categories.

The results in Table 2 indicate that, of the 5,639 forecasts that are supported by alternative data, 1,944 (34%) are based on web traffic data. The next most popular categories are other (23%), followed by point of sale (19%), sentiment (19%), employee (10%), and app usage (8%). The least popular categories are geospatial (5%) and satellite image (3%).

Figure 1 displays two timelines. The first timeline indicates when – for our sample of firms in the DJI – we observe the first analyst report explicitly referencing the use of alternative data from a given alternative data category. The second timeline indicates when we observe the first one hundred analyst reports. The sequence for when we observe the first analyst report is as follows: sentiment (June 11, 2009), web traffic (June 12, 2009), point of sale (August 6, 2009), employee (January 5, 2010), geospatial (January 22, 2010), satellite image (May 3, 2010), other (June 12, 2009) and app usage (July 27, 2010). In other words, within essentially the first year of our sample period, we find that analysts explicitly reference the use of alternative data from all eight categories.

The sequence for when we observe the first one hundred analyst reports is as follows: web traffic (January 19, 2010), point of sale (March 28, 2011), other (August 11, 2011), sentiment (October 10, 2011), satellite image (May 21, 2012), geospatial (June 5, 2012), app usage (September 8, 2014) and employee (August 18, 2015). In other words, by mid-2012, we find that analysts extensively use alternative data from six of the eight categories. Only app usage- and employee data were not widely adopted until mid-2015.

Overall, our evidence suggests that analysts frequently draw from alternative data, at least for the largest and economically most meaningful firms. Our evidence also shows that analysts have been drawing from alternative data, in all their various forms, for a long time.

3. Does the Use of Alternative Data Lead to More Accurate Beliefs?

Analysts presumably discuss the use of alternative data in their reports because the alternative data meaningfully impacted their beliefs about the corresponding company. The impact may be positive, neutral, or negative.

The null hypothesis is that, on average, alternative data do not shift analyst forecasts closer to the actual realizations. One possible reason is that despite prior evidence that alternative data directionally predict performance, it is unclear whether alternative data can be used to consistently develop more accurate *point* estimates, such as what the specific earnings will be at the next annual earnings announcement. In line with this view, some practitioners suggest that the amount of noise in alternative data makes it impossible to uncover signals on a consistent basis (Hope, 2016). Moreover, even if alternative data contained the occasional value-relevant signal, the signal may not be unique. For instance, even if websites compiling employee company ratings contained signals about employee morale, analysts may have been able to procure similar insights from talking to employees on one of their many site visits. Thus, while alternative-data signals may be more convenient to obtain (explaining analysts' adoption of alternative data), in the end, they provide similar insights and therefore do not incrementally improve analysts' forecast accuracy.

Adopting alternative data may even hurt forecast accuracy as acquiring and interpreting alternative data shifts analysts' resources and attention away from other more accurate information channels.

3.1 Empirical Design

To quantify the impact of alternative data on forecast accuracy, we estimate the following regression equation:

$$Acc_{i,f,t} = \eta_{i,f} + \theta_{f,t} + \beta\, I(Alternative\ Data_{i,f,t}) + \gamma`\, Controls + \varepsilon_{i,f,t} \qquad (1)$$

The observations are at the analyst/firm/forecast date level. $Acc_{i,f,t}$ measures analysts' forecast accuracy. The construction of $Acc_{i,f,t}$ follows prior literature (e.g., Clement, 1999; Bradley, Gokkaya, and Liu, 2017; Green, Jame, Markov, and Subasi, 2014; Harford, Jiang, Wang, and Xie, 2019). We first compute $AFE_{i,f,t}$

as the absolute value of the difference between analyst *i*'s annual earnings forecast for firm *f* and the actual reported annual earnings. We then construct $PMAFE_{i,f,t}$ as the difference between $AFE_{i,f,t}$ and $Avg(AFE)_{f,t}$, scaled by $Avg(AFE)_{f,t}$ to reduce heteroskedasticity. $Avg(AFE)_{f,t}$ is the average absolute forecast error across all analysts covering firm *f*. $PMAFE_{i,f,t}$ measures analyst *i*'s forecast accuracy relative to the forecast accuracies of all analysts covering the same firm at the same time. Negative values, or lower forecast errors, indicate above-average performance. Positive values, or higher forecast errors, indicate below-average performance. To facilitate interpretation, $Acc_{i,f,t}$ equals $PMAFE_{i,f,t} \times (-1)$.[11] We provide descriptive statistics regarding *Acc* and other variables we use in this study in Online Appendix Table A3.

Controls include the following analyst characteristics: *Forecast Age*, *Analyst/Firm Experience*, *Analyst Experience, #Firms Covered, Forecast Frequency*, and *Broker Size*. We detail the construction of these variables in Appendix 2. We do not control for firm characteristics as our fixed effects subsume them. Since our final sample comprises 64,036 written reports and earnings forecasts, the number of observations on which we estimate regression equation (1) is 64,036. We double-cluster our standard errors at the analyst- and year-month levels.

We include both analyst-firm ("group"), $\eta_{i,f}$, and firm-year ("period"), $\theta_{f,t}$, fixed effects. Angrist and Pischke (2008) show that our two-way fixed-effects specification is equivalent to the basic difference-in-differences specification. The estimate of *I(Alternative Data)* thus indicates how much more accurate an analyst becomes in the post-adoption period relative to the pre-adoption period compared with analysts covering the same firm over the same period that do not draw from alternative data.

While our estimate does measure the abnormal change in forecast accuracy, the estimate cannot tell us how much of the abnormal change is truly caused by alternative data. The reason is that alternative data adoption is not exogenous. Specifically, alternative data adoption may coincide with an analyst's decision to exert greater effort covering the corresponding firm, which, in turn, leads to improved forecast accuracy ("increased effort channel").

[11] In robustness checks, we base our analysis on $AFE_{i,f,t}$ (Bradley, Gokkaya, and Liu, 2017). As shown in Online Appendix Table A2, the results based on the absolute forecast error are similar to those based on $Acc_{i,f,t}$.

We try to gauge the relevance of the increased effort channel by constructing various measures of analyst effort used in the literature, including the timeliness of forecasts, the number of forecast revisions, and analyst activity during earnings conference calls (Merkley, Michaely, and Pacelli, 2017; Hwang, Liberti, and Sturgess, 2019; Grennan and Michaely, 2020). We then evaluate whether the adoption of alternative data coincides with greater effort. As detailed and tabulated in Online Appendix Table A4, alternative data adoption changes neither the timeliness of forecasts nor the number of forecast revisions. The adoption of alternative data also changes neither the number of questions asked during earnings conference calls, nor the number of words spoken, nor the types of questions asked. The adoption of alternative data coincides with marginally increased conference call attendances.

While we generally fail to find empirical support for the increased effort channel, our tests may lack power. Our point estimate of how much an analyst's forecast accuracy improves after she adopts alternative data should be interpreted with this caveat in mind.

3.2 Evidence

We present our regression results in Table 3. The results reported in column (1) show that the coefficient estimate of *I(Alternative Data)* is 0.214 (*t*-statistic = 6.58). To illustrate the economic significance of this estimate, a 0.214 improvement would move an analyst who is at the median in terms of forecast accuracy to the 62nd percentile.

Another way to gauge the economic significance is to compare the estimate of *I(Alternative Data)* with those of our control variables. For instance, column (1) shows that forecast accuracy increases significantly with the number of years an analyst has been covering a particular firm: the estimate of *Analyst/Firm Experience* is 0.058 (*t*-statistic = 2.81). Comparing the estimate of *I(Alternative Data)* with that of *Analyst/Firm Experience* suggests that the improvement in performance that accompanies the adoption of alternative data is equivalent to having covered the corresponding firm for 3.7 additional years.

As discussed in the previous subsection, endogeneity may cause our point estimate to be upward biased. At the same time, the estimate is so large (and the evidence regarding the increased effort channel so inconclusive) that it does not appear unreasonable to presume that the true effect of alternative data adoption is still greater than zero.[12]

In additional analyses, we replace *I(Alternative Data)* with eight indicator variables, each denoting whether an analyst uses alternative data from a particular alternative data category. We tabulate our findings in column (2). The results show that the adoption of alternative data from six of the eight categories is associated with statistically significant performance improvements. Within those six, the ranking in descending order based on the magnitude of the coefficient estimates is as follows: (1) app usage, (2) sentiment, (3) employee, (4) other, (5) point of sale, and (6) web traffic.

Unlike the adoption of alternative data from the above six categories, our results show that geospatial data and satellite image data are not associated with more accurate earnings forecasts. As shown in Table 2, these are also the two categories that analysts report using least frequently. In 2017, Ernst & Young Global Limited surveyed hedge funds and asked which datasets, in their experience, have been the *least* accurate and *least* insightful.[13] The two datasets that are by far the most frequently mentioned are "geolocation" and "satellite." While the survey conducted by Ernst & Young Global Limited represents a one-time snapshot of investors' opinions, we nevertheless find the overlap between the survey results and our regression results revealing.[14]

4. Where Could the Benefits of Alternative Data Come From?

4.1 Conjectures

We conjecture that alternative data possess three key strengths, which could explain why analysts adopt alternative data and why the adoption appears to lead to more accurate beliefs. First, alternative data provide

[12] Since we have no data on the incremental costs of alternative data and cannot possibly compute the incremental cash flows that the analysts generate as a result of their adoption, we cannot say much regarding the "net present value" of alternative data adoption.
[13] The survey results are also viewable at https://alternativedata.org/stats/.
[14] Through separate analyses tabulated in Online Appendix Table A5-8, we make the same observations when considering alternative data categories separately (as opposed to in one multiple regression). In addition, we find that an analyst's performance improves further if she simultaneously draws from multiple alternative data categories. The source of the alternative data (in-house data-science team versus external data vendor) has no impact on an analyst's performance improvement. The performance improvements strengthen as the analyst gains experience working with alternative data.

instantaneous signals regarding a company's performance. Second, alternative data are compiled by third-party vendors. The fact that managers and firms are not inserted into the data-generating process mitigates misrepresentation concerns. Third, alternative data provide signals at a disaggregated level, such as at the product or branch level. The granularity of alternative data may give investors a more nuanced picture of a company's performance.

To test the relevance of these three possible advantages, we examine whether analysts more frequently adopt alternative data and whether the ensuing performance improvements are stronger (1) when receiving instantaneous signals regarding a company's performance is more important, (2) when traditional data are ambiguous, and (3) when analysts likely lack access to granular data.

Receiving instantaneous signals regarding a company's performance is presumably most useful when there are relatively few company announcements and when the uncertainty regarding a company's performance is high. We thus conjecture that the benefits from alternative data are higher when a firm files relatively few Form 8-Ks, when stock return volatility is high, and when the absolute value of earnings surprises is high.

The fact that managers and firms are removed from the data-generating process becomes particularly relevant when misrepresentation concerns are high and traditional data are ambiguous. As measures for concerns of misrepresentation and the ambiguity of the information environment, we consider earnings restatements (Wilson, 2008) and the absolute value of discretionary accruals (Bhattacharya, Desai, and Venkataraman, 2013).

Finally, the high granularity that alternative data offer is presumably most useful when analysts cannot obtain performance signals at a detailed level through other channels. Private meetings with management are one of the key channels through which analysts and investors can obtain a more nuanced perspective of a company's performance (Green, Jame, Markov, and Subasi, 2014; Soltes, 2014; Soltes and Solomon, 2015; Brown, Call, Clement, and Sharp, 2015; Bengtzen, 2017). Not all analysts are granted private meetings with management, however, putting them at a significant disadvantage. Here, we examine whether alternative data can mitigate the disadvantage.

To measure whether an analyst *i* has preferential access to the management of firm *f*, we consider whether analyst *i* works for a broker that hosts an investor conference in which firm *f* participates. Green, Jame, Markov, and Subasi (2014) argue that broker-hosted investor conferences, which provide select investors opportunities to interact with senior corporate managers, provide insight into whether a particular analyst has preferential access to the firms participating in the conference.

4.2 Evidence - Variation in Analysts' Reliance on Alternative Data

To examine analysts' adoption decisions, we estimate the following probit regression:

$$I(Alternative\ Data_{i,f,t}) = \alpha + \beta` X_{i,f,t} + \delta` Control_{i,f,t} + \varepsilon_{i,f,t} \quad (2)$$

The observations are at the analyst/firm/forecast date level. *I(Alternative Data)* equals one if the analyst issues an earnings forecast that is explicitly supported by alternative data and zero otherwise. *X* includes *Rank(Number of 8-Ks)*, *Rank(Return Volatility)*, *Rank(Earnings Surprise)*, *I(Earnings Restatement)*, *Rank(Discretionary Accruals)*, and *I(Lack of Preferential Access to Management)*. *Number of 8-Ks* is the total number of Form 8-Ks filed during the previous annual forecast period. *Return Volatility* is the standard deviation of daily stock returns in the previous annual forecast period. *Earnings Surprise* refers to the most recent earnings surprise in the previous forecast period, measured using quarterly diluted earnings per share, excluding extraordinary items, and applying a seasonal random walk (Livnat and Mendenhall, 2006). *I(Earnings Restatement)* equals one if the corresponding firm has had to restate its financial accounts. We compute *Discretionary Accruals* as in Kothari, Leone, and Wasley (2005) based on the most recent annual financial statement announcement. To facilitate a comparison of the coefficient estimates, we convert *Number of 8-Ks, Return Volatility, Earnings Surprise,* and *Discretionary Accruals* into quintile rank variables, ranging from one if the corresponding realization is in the bottom quintile of its distribution to five if the corresponding realization is in the top quintile. *I(Lack of Preferential Access to Management)* equals one if the corresponding firm did not participate in a conference hosted by the corresponding analyst's broker over the previous year.

Controls include the following analyst and firm characteristics: *I(In-House Data Science Team)*, $\sum$ *Colleagues* $_{Alternative\ Data}$, *Analyst/Firm Experience*, *Analyst Experience*, *#Firms Covered*, *Forecast Frequency*, *Broker Size*, *Size*, *M/B*, and *Momentum*. We detail the construction of these variables in Appendix 2. We double-cluster our standard errors at the analyst- and year-month levels.

Table 4 presents the results. The estimates of *Rank(Return Volatility)*, *Rank(Earnings Surprise)*, and *I(Earnings Restatement)* are all strongly positive. The estimates of *Rank(Discretionary Accruals)* and *I(Lack of Preferential Access to Management)* are mildly positive. The estimate of *Rank(Number of 8-Ks)* is strongly negative. All these estimates are consistent with our predictions.

To illustrate the economic significance, our estimates in column (3), which considers all predictors jointly in one regression equation, suggest that a one quintile decrease in the number of Form 8-Ks increases the likelihood of alternative data adoption by 1.4% (*z*-value = -3.46). The corresponding increases in likelihood tied to rises in stock return volatility, earnings surprises and discretionary accruals are 1.5% (*z*-value = 4.09), 0.7% (*z*-value = 2.57) and 0.6% (*z*-value = 1.77), respectively. The increase in likelihood tied to an earnings restatement and having no preferential access to management are 4.4% (*z*-value = 3.14) and 2.1% (*z*-value = 1.73). To put these numbers in perspective, the average fraction of forecasts explicitly supported by alternative data is 9%. The implied increases in the likelihood of alternative data adoption are thus sizeable.

4.3 Evidence – Variation in the Usefulness of Alternative Data

We next examine whether the features that increase alternative data adoption also improve analysts' forecast accuracy. We re-estimate our main regression equation (1) separately for two subsamples and compare the coefficient estimates of *I(Alternative Data)*. Specifically, we separate observations by whether the number of Form 8-Ks is in the top versus the bottom quintile of its distribution. We then estimate regression equation (1) separately for each of the two subsets. Similarly, we separate observations by whether the stock return volatility, the earnings surprise, and the discretionary accruals are in the top- versus the bottom quintile. We also separate observations into those for which the corresponding firm, as of the corresponding earnings report date, has had to restate its financial accounts and those for which there has been no earnings restatement. Finally, we separate observations by whether, over the past year, the corresponding analyst's broker did not host a conference in which the corresponding firm was a participant.

We report our results in Table 5. Consistent with our predictions, we find that the estimate of *I(Alternative Data)* is substantially larger when the corresponding firm issues relatively few Form 8-Ks, when earnings surprises are large, when the firm has had to restate its financial accounts and when the corresponding firm has high levels of discretionary accruals. We find that the performance improvements ensuing alternative data adoption are marginally larger when stock return volatility is high and when the corresponding analyst has no investor-conference tie.

Overall, our evidence suggests that alternative data can partially address some of the shortcomings of traditional data and help financial market participants form more accurate beliefs. Alternative data can thus help make markets more efficient.

Our evidence also suggests that the incremental benefit of alternative data varies by firm and analyst characteristics. This variation helps explain why a corresponding analyst may adopt alternative data for some of the firms she covers but not for others and why for a corresponding firm, some analysts adopt alternative data but not others.

5. Alternative Data and Stock Market Reactions

Our final analysis tests whether the stock market recognizes the seeming usefulness of alternative data. To tackle this question, we estimate separate versions of the following regression equation that include, in turn, changes in earnings forecasts, changes in target prices, and changes in overall stock recommendations:

$$\begin{aligned} CAR_{i,f,t} = {} & \eta_{i,f} + \theta_{f,t} + \beta_1\, I(Alternative\ Data_{i,f,t}) + \beta_2\, \Delta_{i,f,t} \\ & + \beta_3\, I(Alternative\ Data_{i,f,t}) \times \Delta_{i,f,t} + \gamma`\, Controls + \varepsilon_{i,f,t}. \end{aligned} \tag{3}$$

The observations are at the analyst/firm/forecast date level. Following Green, Jame, Markov, and Subasi (2014), we delete observations within two trading days of a quarterly earnings announcement to avoid the impact of confounding events. $CAR_{i,f,t}$ is the cumulative market-adjusted return over days [0,+1], where day 0 is the report date of the earnings forecast, the target price, or the recommendation. If the report date falls on a non-trading day, Day 0 is the ensuing trading day. $\Delta_{i,f,t}$ is either the percentage change in the earnings forecast, the percentage change in the target price, or the change in the "recommendation score." The percentage change in the earnings forecast is the difference between the current annual earnings forecast and the previous annual earnings forecast issued by analyst *i* for firm *f* scaled by the absolute value of the previous annual earnings forecast. The percentage change in the target price is the difference between the current target price and the previous target price scaled by the previous target price. To compute the change in the recommendation score, we first convert recommendations to numerical scores: 1 for sell recommendations, 2 for hold recommendations, and 3 for buy recommendations. The change in the recommendation score is the difference between the current score and the previous score.

Controls include the following analyst and firm characteristics, all described in Appendix 2: *Forecast Age, Analyst/Firm Experience, Analyst Experience, #Firms Covered, Forecast Frequency, Broker Size, Size, M/B,* and *Momentum.* As before, we double-cluster our standard errors at the analyst- and year-month levels.

Asquith, Mikhail, and Au (2005), Green, Jame, Markov, and Subasi (2014), and Bradley, Gokkaya, and Liu (2017), among others, detect a positive association between stock-price changes and changes in earnings forecasts, changes in target prices, and changes in overall stock recommendations, respectively. That is, prior literature finds that the estimate of Δ is positive. Here, we test whether the positive association becomes stronger when analysts draw from alternative data. That is, we test whether the estimate of the interaction term, *I(Alternative Data)* $\times$ Δ, is positive and significant.

We report the results in Table 6. For columns (1) and (2), we consider the market's reaction to changes in earnings forecasts without and with fixed effects, respectively. For columns (3) and (4), we consider the market's reaction to changes in target prices. For columns (5) and (6), we consider the market's reaction to changes in overall stock recommendations.

For all specifications, we find that the estimate of the interaction term is positive, statistically significant, and economically meaningful. Our results thus suggest that the market perceives analyst research as more informative when an analyst chooses to incorporate alternative data.

To illustrate the economic significance of this result, our estimate suggests that, on average, when an analyst does not incorporate alternative data and revises her earnings forecast upward by 10%, the corresponding stock price increases by 0.4%. This estimate is similar to that found in prior literature. For instance, Bradley, Gokkaya, and Liu (2017) find that a 10% upward revision in an analyst's earnings forecast is accompanied by a 0.6% stock-price increase. In comparison, when an analyst incorporates alternative data and revises her earnings forecast upward by 10%, our estimate suggests that the corresponding stock price increases by 1.2%. In other words, the change in the stock price is roughly three times larger when an analyst incorporates alternative data.

Similarly, the results reported in columns (3) through (6) suggest that when an analyst incorporates alternative data, the stock market responds twice as strongly to changes in target prices or stock recommendation levels as when an analyst does not incorporate alternative data.

In additional analyses, we consider cumulative market-adjusted return over the ensuing week, month, and year. As shown in Online Appendix A9, we do not find systematic return continuations or reversals in the medium- and long-run.

6. Conclusion

Our study documents how widely analysts use alternative data, in what form, and what types of situations. We also provide suggestive evidence on the performance implications of alternative data adoption. The high cost and the associated unequal access to alternative data are frequently broached as threats to equality and level playing fields (Hilbert, 2016; O'Neil, 2016). Our evidence suggests that, at least in the investment domain, this threat is mitigated by the presence of important information intermediaries.

Future work may consider other possible challenges that come from the emergence of big data. One possible threat we can think of is that market participants eventually become overly excited about the prospects of big data and overweigh the common, quantitative signals provided by big data. The popularization of big data may thus prevent useful private information from entering market prices (Da and Huang, 2020). In addition, if big data provide more information about cash flows in the near term than cash flows in the distant future, the emergence of big data may shift market participants' efforts away from predicting cash flows in the distant future and, in that regard, make markets less efficient (Dessaint, Fresard, and Foucault, 2021). Examining these and related possible implications should be a fruitful avenue for future research.

References

Ambert K, Beaune S, Chaibi A, Briard L, Bhattacharjee A, Bharadwaj V, Sumanth K, Crowe K (2016) French hospital uses trusted analytics platform to predict emergency department visits and hospital admissions. Intel White Paper.

Agarwal S, Qian W, Zou X (2021) Disaggregated sales and stock returns. *Management Sci*. 67(11): 7167-7183.

Angrist JD, Pischke JS (2008) *Mostly harmless econometrics* (Princeton University Press, Princeton, NJ).

Asquith P, Mikhail MB, Au AS (2005) Information content of equity analyst reports. *J. Financial Econom.* 75(2): 245-282.

Bartov E, Faurel L, Mohanram PS (2018) Can Twitter help predict firm-level earnings and stock returns? *Accounting Rev*. 93(3): 25-57.

Bengtzen M (2017) Private investor meetings in public firms: the case for increasing transparency. *Fordham J. Corporate & Financial Law* 22 (1): 33-132.

Bhattacharya N, Desai H, Venkataraman K (2013) Does earnings quality affect information asymmetry? Evidence from trading costs. *Contemp. Accounting Res.* 30(2): 482-516.

Bradley D, Gokkaya S, Liu X (2017) Before an analyst becomes an analyst: does industry experience matter? *J. Finance* 72(2): 751-792.

Brown LD, Call AC, Clement MB, Sharp NY (2015) Inside the "black box" of sell-side financial analysts. *J. Accounting Res.* 53(1): 1-47.

Clement MB (1999) Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *J. Accounting Econom*. 27(3): 285-303.

Da Z, Huang X (2020) Harnessing the wisdom of crowds. *Management Sci.* 66(5): 1847-1867.

Dannemiller D, Kataria R (2017) Alternative data for investment decisions: today's innovation could be tomorrow's requirement. Deloitte Center for Financial Services Report.

Dessaint O, Foucault T, Frésard L (2021) Does alternative data improve financial forecasting? The horizon effect. Working paper, INSEAD, Fontainebleau, France.

Diamond DW, Verrecchia RE (1991) Disclosure, liquidity, and the cost of capital. *J. Finance* 46(4): 1325-1359.

Dichev ID, Qian J (2022) The benefits of transaction-level data: The case of NielsenIQ scanner data. *J. Accounting Econom.* Forthcoming.

Easley D, O'Hara M (1987) Price, trade size, and information in securities markets. *J. Financial Econom.* 19(1): 69-90.

Glosten LR, Milgrom PR (1985) Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *J. Financial Econom.* 14(1): 71-100.

Green TC, Huang R, Wen Q, Zhou D (2019) Crowdsourced employer reviews and stock returns. *J. Financial Econom*. 134(1): 236-251.

Green TC, Jame R, Markov S, Subasi M (2014) Access to management and the informativeness of analyst research. *J. Financial Econom.* 114(2): 239-255.

Grennan J, Michaely R (2020) Artificial intelligence and high-skilled work: Evidence from analysts. Working paper, Duke University, Durham, NC.

Gupta T, Leung E, Roscovan V (2022) Consumer spending and the cross section of stock returns. *J. Portfolio Management*. Forthcoming.

Harford J, Jiang F, Wang R, Xie F (2019) Analyst career concerns, effort allocation, and firms' information environment. *Rev. Financial Stud*. 32(6): 2179-2224.

Hilbert M (2016) Big data for development: A review of promises and challenges. *Development Policy Rev.* 34(1): 135-174.

Hoberg G, Moon SK (2017) Offshore activities and financial vs. operational hedging. *J. Financial Econom*. 125(2): 217-244.

Hoberg G, Moon SK (2019) The offshoring return premium. *Management Sci.* 65(6): 2876-2899.

Hong H, Kubik JD (2003) Analyzing the analysts: career concerns and biased earnings forecasts. *J. Finance* 58(1): 313-351.

Hope B (2016) Wall Street's love: data, data, data. *Wall Street Journal*. Retrieved from https://www.wsj.com

Huang J (2018) The customer knows best: the investment value of consumer opinions. *J. Financial Econom.* 128(1): 164-182.

Huang K, Li M, Markov S (2020) What do employees know? Evidence from a social media platform. *Accounting Rev*. 95(2): 199-226.

Hwang BH, Liberti JM, Sturgess J (2019) Information sharing and spillovers: evidence from financial analysts. *Management Sci.* 65(8): 3624-3636.

Kothari SP, Leone AJ, Wasley CE (2005) Performance matched discretionary accrual measures. *J. Accounting Res*. 39(1): 163-197.

Livnat J, Mendenhall RR (2006) Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. *J. Accounting Res*. 44(1): 77-205.

Mayer-Schönberger V, Cukier K (2013) *Big data: A revolution that will transform how we live, work, and think* (Houghton Mifflin Harcourt, Boston, MA).

McMahon D, Chu K (2012) Clampdown in China on corporate sleuthing. *Wall Street Journal*. Retrieved from https://www.wsj.com

Merkley K, Michaely R, Pacelli J (2017) Does the scope of the sell-side analyst industry matter? An examination of bias, accuracy, and information content of analyst reports. *J. Finance* 72(3): 1285-1334.

Mukherjee A, Panayotov G, Shon J (2021) Eye in the sky: private satellites and government macro data. *J. Financial Econom.* 141(1): 234-254.

O'Neil C (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy* (Broadway Books, New York, NY).

PRNewswire (2022) Global alternative data market report 2022-2030: Rising demand from hedge funds is expected to propel growth. Retrieved from https://www.prnewswire.com/

Ram A, Wigglesworth R (2017) When silicon valley came to Wall Street. *Financial Times*. Retrieved from https://www.ft.com

Savi R, Shen J, Betts B, MacCartney (2015) Finding big alpha in big data. BlackRock Report.

Sheng J (2022) Asset pricing in the information age: employee expectations and stock returns. Working paper, University of California - Irvine, Irvine, CA.

Solomon D, Soltes E (2015) What are we meeting for? The consequences of private meetings with investors. *J. Law & Econom*. 58(2): 325-355.

Soltes E (2014) Private interaction between firm management and sell-side analysts. *J. Accounting Res*. 52(1): 245-272.

Watts W (2019) The explosion of 'alternative' data gives regular investors access to tools previously employed only by hedge funds. *MarketWatch*. Retrieved from https://www.marketwatch.com

Whyte A (2017) The world's best research firm. *Institutional Investor*. Retrieved from https://www.institutionalinvestor.com

Wigglesworth R (2016) Investors mine big data for cutting-edge strategies. *Financial Times*. Retrieved from https://www.ft.com

Wigglesworth R (2020) Stockpickers turn to big data to arrest decline. *Financial Times*. Retrieved from https://www.ft.com

Wilson WM (2008) An empirical analysis of the decline in the information content of earnings following restatements. *Accounting Rev*. 83(2): 519-548.

Appendix 1
List of Alternative Data Categories and Keywords

Column (1) reports our eight alternative data categories. Column (2) reports our definitions of each category. Column (3) reports the list of keywords that we deem as pointing to the use of alternative data from a particular category. In our search, we include common variations of the keywords, such as singular and plural, past and present tense, uppercase, and lowercase. Keywords ending with "*" are names of alternative-data vendors. Column (4) provides excerpts from analyst reports describing the use of alternative data from the corresponding category.

Category (1)	Definition (2)	Keywords (3)		Example from Analyst Report (4)
App Usage	These data track the number of downloads, the number of active users, and the time spent on mobile apps.	active user App Annie* AppData* Jiguang* QuestMobile* Sensor Tower*	SimilarWeb* TalkingData*	The UBS Evidence Lab analyzed App data that provides wait times for the 24 Shanghai Disneyland attractions that have wait times associated with them. Our analysis covers the thirteen-week period from November 6, 2016 through January 29, 2017. [Issued by UBS on 04/06/17 for WALT DISNEY CO]
Sentiment	These data include social-media feeds and news flow that help gauge consumer sentiment on products and services.	brand sentiment CMS Data* consumer sentiment customer rating customer review customer satisfaction rating customer satisfaction trend facebook analysis facebook data facebook like facebook likes facebook post facebook track facebook user guest sentiment instagram data instagram engagement instagram follower Internet World Stats* Investing Analytics* Medicare Plan Finder* Merchant Centric* net sentiment NetBase*	online customer review online review Prosper Insights* ratings on tripadvisor review analytics scoring released by cms sentiment analysis sentiment data social media analysis social media engagement social media follower star(s) rating tracking on twitter tripadvisor ratings twitter analysis twitter data twitter purchase intent twitter sentiment web analytics web mining web scraping yelp	In this report, we introduce our proprietary consumer sentiment analysis, using information from Merchant Centric, a company that works with multi-location brands across consumer and service industries to "help them manage and learn from guests' online feedback." For our purposes, Merchant Centric tracks location-specific, user-generated reviews across multiple social media platforms. The reviews are user-generated and tied to a specific location, and then sourced from the following social media sites: Facebook, Google, Yelp,Trip Advisor, Superpages, and CitySearch. Our specific Jefferies data set utilizes reviews on a representative sample of some 500 McDonald's locations across the country, as well as reviews for just over 2,400 Bojangles, Burger King, Del Taco, Dunkin' Donuts, Jack in the Box, Sonic, Taco Bell, and Wendy's units located within the same zip code. We have chosen to exclude independent/local operators, and focus on a sample of national and regional competitors. [Issued by Jefferies on 12/05/17 for MCDONALD'S CORP]

Appendix 1. Continued.

Category (1)	Definition (2)	Keywords (3)		Example from Analyst Report (4)
Employee	These data include job postings to evaluate corporate strategy, industry growth rates, and demand for specific job skills. These data also include management- and employee sentiment extracted from statements in earnings conference calls and sites such as Glassdoor, among others.	earnings call transcript indeed.com job posting job trend mining of earnings calls online hiring	text analytics transcriptlytics web analytics web mining web scraping	We do track Apple's overall job postings and have seen a notable increase over the past 4-5 months in the number of engineering positions for Siri and ML, with a total of 205 specific mentions of "Siri," "deep learning," "computer vision," "natural language processing (NLP) or "machine learning" in March job postings up from 64 mentions back in November. [Issued by GUGGENHEIM on 04/10/18 for APPLE]
Geospatial	These data include store location data to analyze the local competition, often overlayed with local income data and other demographic information to assess demand.	branch network model branch rationalization tool demographic analysis Foursquare* geopspatial	market quality analysis store overlap within […] drive within […] miles	We utilized the Alpha-Wise Branch Network Model to preview markets where we think JPM will likely invest. Seven factors drove our rankings, including wealth, income and population growth, competitive intensity, and small business opportunities. We calculate this as average deposits per branch. Our view is that areas with more deposits per branch are attractive for two reasons: 1) it's indicative of concentrated wealth and 2) it could suggest the area is underserved by low branch count. [Issued by MORGAN STANLEY on 02/21/18 for J.P.MORGAN]
Point of Sale	These data include merchant level transaction data (e.g. retailer, airline, service provider), product level purchase data (e.g. food, beverages, electronics) and pricing data.	1010Data* airbnb + listing compared online prices discount tracker financial rate monitor First Data SpendTrend* footlocker.com footwear scrapes hotel tracker hotel tracking listing monitor MasterCard Advisors* Nielson* online price survey online pricing study our proprietary datasets Point-Of-Sale*	price comparisons price intelligence price monitoring price observations pricing monitor pricing study pricing tracker property listing Sg2* spend tracker Standard Media Index* SuperData* vehicle listing web analytics web mining web scraping zillow	CS Proprietary Home Pricing Tracker (median home price trends on an individual store basis across entire store base) shows similar trends in HD/LOW markets. [Issued by CREDIT SUISSE on 02/19/16 for HOME DEPOT]

Appendix 1. Continued.

Category (1)	Definition (2)	Keywords (3)		Example from Analyst Report (4)
Satellite Image	Satellite images can be used to track consumer traffic as well as production activities at mines, construction sites, plants, and oil and gas companies, among others.	Orbital Insight* parking lot fill rate parking lot traffic proprietary satellite data remote sensing	Remote Sensing Metrics* RS Metrics* satellite analysis satellite image traffic analysis	The satellite analysis points to a y/y Q1 parking lot fill rate change of +0.4%, however, the y/y change in fill rate became progressively worse over the quarter. Based on this data and the headwinds faced in Q1 from cash for appliances and weather, we feel comfortable with our Q1E comp of +1%. [Issued by PIPER SANDLER on 05/12/11 for HOME DEPOT]
Web Traffic	These data track what users search for in the Internet and how frequently/for how long users visit given websites.	baidu analysis baidu data baidu search data baidu search index baidu search volume ComScore* daily traffic google search analysis google search trend google trend google-searched iphone monitor	iphone tracker Scrapehero* search interest search trend search volume smartphone tracker Thinknum* traffic analysis traffic monitor web hit activity web search	The AlphaWise Smartphone Tracker has been developed by Morgan Stanley's AlphaWise using multi-country web search analysis using Google Trends. The approach accounts for different search criteria in multiple countries, as well as the differential between search and sales data seasonality, where appropriate. The in-sample period consists of 2008-2011 for Apple and 2010-2012 for Samsung Galaxy. [Issued by MORGAN STANLEY: on 09/18/13 for APPLE]
Other	These include alternative data which do not fit cleanly to any of the categories above. Examples are clipper data and macro demand data.	BuildFax* climatology ClipperData* Collateral Verifications* Dodge* Drillinginfo* Dun & Bradstreet* Edmunds* Entgroup* EPFR* evidence lab macro Flightglobal* formulary coverage home improvement tracker IFI Claims* Innovata*	lower end spending m2m macro-to-micro network traffic lab nowcast One Click Retail* Ookla* OpenSignal* Root Metrics* Rystad* STR data* wait time monitor Wards Automotive* weather monitor	The most effective way to measure the integrated advantage is our proprietary use of ClipperData, which can track the shipper ID of barrels loading onto vessels in the Gulf of Mexico. For this analysis, we do not isolate the loadings to export barrels, but look at Jones Act activity as well, as the ability to move its crude production anywhere is the proper reflection of the business model's advantage, in our view. [Issued by WOLFE RESEARCH on 09/28/18 for EXXON MOBIL CORP]

Appendix 2
Variable Description

Variables	Definition
Acc	We first calculate the proportional mean absolute forecast error, *PMAFE*, as the difference between the absolute forecast error of an analyst and the average absolute forecast error across all analysts, scaled by the average absolute forecast error. Since negative (positive) values of *PMAFE* indicate above (below) average performance, *Acc* is defined as *PMAFE* × (-1).
I(Alternative Data)	An indicator variable that equals one if the corresponding analyst issues an earnings forecast explicitly supported by alternative data and zero otherwise.
Forecast Age	The logarithm of one plus the number of calendar days between the forecast date and the corresponding I/B/E/S report date of the actual earnings.
Analyst/Firm Experience	The number of years since the corresponding analyst first issued a forecast for the corresponding firm.
Analyst Experience	The number of years since the corresponding analyst first issued a forecast for any firm in the IBES database.
#Firms Covered	The logarithm of one plus the number of firms the corresponding analyst covers in the corresponding year.
Forecast Frequency	The logarithm of one plus the number of forecasts made by the corresponding analyst in the corresponding year.
Broker Size	The number of analysts working at the corresponding analyst's broker in the year of the forecast.
Number of 8-Ks	The total number of Form 8-Ks filed during the previous annual forecast period.
Return Volatility	The standard deviation of daily stock returns during the previous annual forecast period.
Earnings Surprise	Most recent earnings surprise in the previous forecast period, which is measured using quarterly diluted earnings per share, excluding extraordinary items, and applying a seasonal random walk (Livnat and Mendenhall, 2006).
I(Earnings Restatement)	An indicator variable that equals one if the firm has issued restatements in the past.
Discretionary Accruals	We calculate discretionary accruals based on the Modified Jones model matched to another from the same industry and year with the closest ROA (Kothari, Leone, and Wasley, 2005).
I(Lack of Preferential Access to Management)	An indicator variable that equals one if the corresponding firm did not participate in a conference hosted by the corresponding analyst's broker over the previous year.
I(In-House Data Science Team)	An indicator variable that equals one if the corresponding analyst works for a broker that has an in-house data science team.
$\sum Colleagues_{Alternative\ Data}$	The number of analysts that rely on alternative data and work for the same broker as the corresponding analyst.
Size	The market capitalization of the corresponding firm at the end of the previous fiscal year in billions.

Appendix 2. Continued.

M/B	The market value of equity divided by the book value of equity at the end of the previous fiscal year.
Momentum	Buy-and-hold return of the corresponding stock over the previous six months.
Cumulative Abnormal Returns	Two-day cumulative abnormal return CAR[0,1] benchmarked against the CRSP value-weighted index return. Day 0 is the announcement date of the forecast revision or the recommendation change.
Earnings Forecast Change	Difference between current and previous earnings forecast issued by the same analyst for the same firm, scaled by the absolute value of the previous forecast.
Target Price Change	Difference between the current and the previous target price issued by the same analyst for the same firm, scaled by the absolute value of the previous target price.
Recommendation Change	Difference between the current and the previous recommendation issued by the same analyst for the same firm. Recommendations are converted to a 1 when they represent sell recommendations, a 2 for hold recommendations, and a 3 for buy recommendations.

Figure 1

This figure displays two timelines, indicating when – for our sample of firms in the Dow Jones Industrial Average Index – we observe the first analyst report, or, the first one hundred analyst reports, explicitly referencing the use of alternative data from a particular alternative data category. We describe our alternative data categories in Section 2.3.

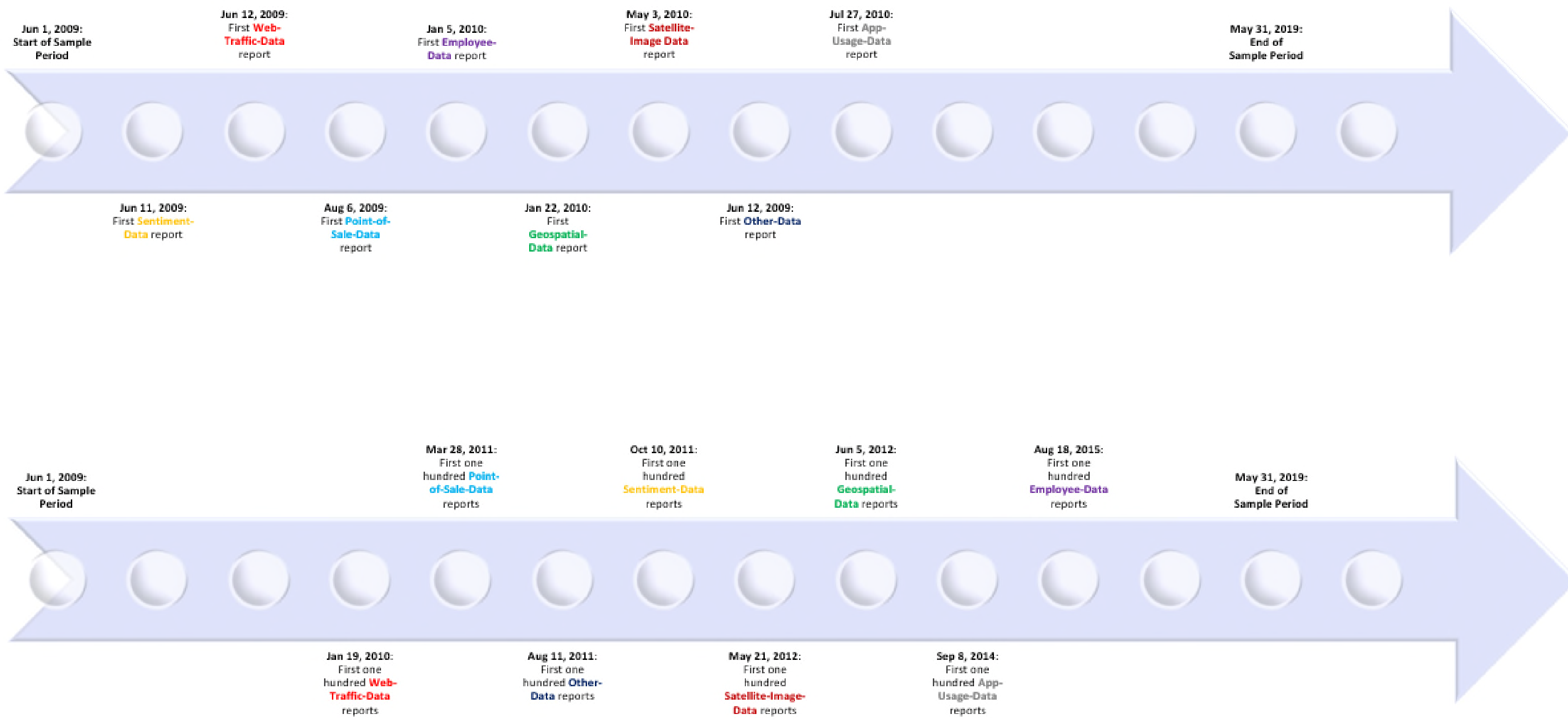

Table 1
Number and Fraction of Analyst Forecasts Explicitly Supported by Alternative Data

In this table, we present the numbers and the fractions of analyst forecasts explicitly supported (not explicitly supported) by alternative data. Our sample contains all Dow Jones Industrial Average Index firms from June 1, 2009, through May 31, 2019. We combine the years 2009 and 2010 and the years 2018 and 2019 as we have only partial data for the years 2009 and 2019.

	Number of Forecasts …		Fraction of Forecasts…
	… Explicitly Supported by Alternative Data	… Not Explicitly Supported by Alternative Data	… Explicitly Supported by Alternative Data
Panel A: By Year			
2009/2010	515	7,634	6%
2011	615	6,239	9%
2012	488	6,769	7%
2013	490	6,348	7%
2014	497	6,058	8%
2015	694	5,998	10%
2016	729	5,691	11%
2017	659	5,444	11%
2018/2019	952	8,216	10%
2009 - 2019	5,639	58,397	9%
Panel B: By Industry Sector			
Energy	40	2,512	2%
Materials	29	2,614	1%
Industrials	580	8,398	6%
Consumer Discretionary	443	4,194	10%
Consumer Staples	841	7,199	10%
Health Care	661	7,487	8%
Financials	103	8,082	1%
Information Technology	2,513	13,597	16%
Communication Services	429	4,314	9%

Table 2
Number of Analyst Forecasts Explicitly Supported by Data from a Particular Category

In this table, we present the numbers of analyst forecasts explicitly supported by data from a particular alternative data category. We describe our alternative data categories in Section 2.3. Since a given analyst report may draw from multiple alternative data categories, the sum of the number of forecasts in Table 2 exceeds the total number of forecasts explicitly supported by alternative data reported in Table 1; the fractions do not add up to 100% for the same reason.

Alternative Data Category	Number [Fractions] of Forecasts Explicitly Supported by Alternative Data
App Usage	476 [8%]
Employee	543 [10%]
Geospatial	257 [5%]
Other	1,322 [23%]
Point of Sale	1,080 [19%]
Satellite Image	171 [3%]
Sentiment	1,062 [19%]
Web Traffic	1,944 [34%]

Table 3
Alternative Data and Forecast Accuracy

This table reports coefficient estimates from regressions of forecast accuracy on whether an analyst explicitly references the use of alternative data in her corresponding written report. The observations are at the analyst/firm/forecast date level. To construct the dependent variable, *Acc*, we first compute *PMAFE* as the difference between the absolute forecast error of an analyst and the average absolute forecast error across all analysts, scaled by the average absolute forecast error. Since negative (positive) values of *PMAFE* indicate above (below) average performance, we define *Acc* as *PMAFE* × (-1). *I(Alternative Data)* equals one if the corresponding analyst's earnings forecast is explicitly supported by alternative data as described in Section 2.2 and zero otherwise. *I(Category = X)* equals one if the corresponding analyst explicitly references the use of alternative data from alternative data category *X*. We define all remaining variables in Appendix 2**.** We report *t*-statistics in parentheses. We double-cluster our standard errors at the analyst- and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)
I(Alternative Data)	0.214***	
	(6.58)	
I(Category = App Usage)		0.307***
		(4.08)
I(Category = Sentiment)		0.220***
		(2.89)
I(Category = Employee)		0.209***
		(3.15)
I(Category = Geospatial)		-0.015
		(-0.14)
I(Category = Point of Sale)		0.182***
		(3.53)
I(Category = Satellite Image)		0.036
		(0.36)
I(Category = Web Traffic)		0.146**
		(2.31)
I(Category = Others)		0.183***
		(3.94)
Forecast Age	-0.246***	-0.245***
	(-12.37)	(-12.39)
Analyst/Firm Experience	0.058***	0.058***
	(2.81)	(2.76)
Analyst Experience	0.061	0.060
	(1.11)	(1.10)
#Firms Covered	0.039	0.040
	(0.77)	(0.79)
Forecast Frequency	0.028	0.028
	(1.02)	(1.00)
Broker Size	-0.001	-0.001
	(-1.03)	(-1.09)
Analyst-Firm Fixed Effects	Yes	Yes
Firm-Year Fixed Effects	Yes	Yes
N	64,036	64,036
Adjusted R^2	0.231	0.232

Table 4
Variation in the Use of Alternative Data

This table reports coefficient estimates from probit regressions of alternative data adoption on various analyst- and firm characteristics. The observations are at the analyst/firm/forecast date level. The dependent variable equals one if the analyst's earnings forecast is explicitly supported by alternative data and zero otherwise. *Number of 8-Ks* is the total number of Form 8-Ks filed during the previous annual forecast period. *Return Volatility* is the standard deviation of daily stock returns during the previous annual forecast period. *Earnings Surprise* is measured as in Livnat and Mendenhall (2006). *I(Earnings Restatement)* equals one if the corresponding firm has had to restate its financial accounts. We compute *Discretionary Accruals* as in Kothari, Leone, and Wasley (2005) based on the most recent annual financial statement announcement. *I(Lack of Preferential Access to Management)* equals one if the corresponding firm did not participate in a conference hosted by the corresponding analyst's broker over the previous year. To facilitate a comparison of the coefficient estimates, we convert *Number of 8-Ks, Return Volatility, Earnings Surprise,* and *Discretionary Accruals* into quintile rank variables, ranging from one if the corresponding realization is in the bottom quintile of its distribution to five if the corresponding realization is in the top quintile. We define all remaining variables in Appendix 2**.** We report *z*-statistics in parentheses. We double-cluster our standard errors at the analyst- and the year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)	(2)	(3)
Rank(Number of 8-Ks)	-0.099***		-0.105***
	(-3.29)		(-3.46)
Rank(Return Volatility)	0.098***		0.115***
	(3.51)		(4.09)
Rank(Earnings Surprise)	0.055***		0.054***
	(2.82)		(2.57)
I(Earnings Restatement)	0.329***		0.308***
	(3.38)		(3.14)
Rank(Discretionary Accruals)	0.043*		0.043*
	(1.77)		(1.77)
I(Lack of Preferential Access to Management)	0.044		0.171*
	(0.47)		(1.73)
I(In-House Data Science Team)		0.413***	0.437***
		(2.89)	(3.31)
$\sum$ *Colleagues* $_{\textit{Alternative Data}}$		0.038**	0.048***
		(2.04)	(2.93)
Analyst/Firm Experience	-0.009	-0.010	-0.006
	(-1.21)	(-1.34)	(-0.79)
Analyst Experience	0.008	0.001	0.003
	(1.08)	(0.19)	(0.43)
#Firms Covered	0.097	0.088	0.154
	(0.68)	(0.63)	(1.15)
Forecast Frequency	0.054	-0.010	-0.013
	(0.78)	(-0.15)	(-0.19)
Broker Size	0.001	-0.000	0.001
	(1.55)	(-0.06)	(0.72)

Table 4. Continued.

	(1)	(2)	(3)
Size	0.237***	0.306***	0.208***
	(3.28)	(3.94)	(3.35)
M/B	0.014	0.010	0.012
	(1.54)	(1.05)	(1.21)
Momentum	0.089	0.572***	0.193
	(0.66)	(2.88)	(1.37)
Analyst-Firm Fixed Effects	No	No	No
Firm-Year Fixed Effects	No	No	No
N	64,036	64,036	64,036
Pseudo R^2	0.087	0.076	0.120

Table 5
Variation in the Usefulness of Alternative Data

This table reports results from repeating the analysis tabulated in column (1) of Table 3, but we now conduct the analysis separately on observations for which we predict alternative data are more advantageous (column (1)) and observations for which alternative data are less advantageous (column (2)). In Panels A, B, C, and E, we separately consider observations in the top and the bottom quintile with regards to *Number of 8-Ks*, *Return Volatility*, *Earnings Surprise*, and *Discretionary Accruals*, respectively. In Panel D, we separate observations by whether the corresponding firm has had to restate its financial accounts or not. In Panel F, we separate observations by whether, over the previous year, the corresponding firm participated in a conference hosted by the corresponding analyst's broker or not. We report *t*-statistics in parentheses. We double-cluster our standard errors at the analyst- and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	Alternative Data …		
	… More Advantageous (1)	… Less Advantageous (2)	*Wald*-Test Statistics of Coefficient Equality
Panel A: Number of 8-Ks ("Bottom Quintile" versus "Top Quintile")			
I(Alternative Data)	0.381***	0.198***	2.757*
	(4.57)	(2.68)	
N	12,638	12,567	
Panel B: Return Volatility ("Top Quintile" versus "Bottom Quintile")			
I(Alternative Data)	0.272***	0.238***	0.152
	(4.65)	(3.35)	
N	13,101	12,742	
Panel C: Earnings Surprise ("Top Quintile" versus "Bottom Quintile")			
I(Alternative Data)	0.389***	0.099*	8.625***
	(4.45)	(1.93)	
N	12,687	12,777	
Panel D: Earnings Restatement ("Yes" versus "No")			
I(Alternative Data)	0.318***	0.120***	9.302***
	(5.74)	(3.42)	
N	20,477	43,559	
Panel E: Discretionary Accruals ("Top Quintile" versus "Bottom Quintile")			
I(Alternative Data)	0.373***	0.154**	2.714*
	(3.34)	(2.05)	
N	12,728	12,843	
Panel F: Preferential Access to Management ("No" versus "Yes")			
I(Alternative Data)	0.230***	0.151***	1.730
	(5.56)	(3.14)	
N	48,125	15,911	

Table 6. Alternative Data and Stock Market Reactions

This table reports coefficient estimates from regressions of cumulative abnormal returns on changes in analyst forecasts. The observations are at the analyst/firm/forecast date level. We remove observations that coincide with quarterly earnings announcements. The dependent variable is the percentage cumulative market-adjusted return in the first two trading days of the forecast change. *I(Alternative Data)* is an indicator variable, which equals one if the corresponding analyst's forecast is explicitly supported by alternative data and zero otherwise. In columns (1) and (2), *Δ* is the percentage change in the earnings forecast. In columns (3) and (4), *Δ* is the percentage change in the target price. In columns (5) and (6), we convert recommendations to numerical scores (1 for sell-, 2 for hold-, and 3 for buy recommendations); *Δ* is the change in the numerical score. We define all remaining variables in Appendix 2. "Firm Characteristics Controls" include *Size*, *M/B*, and *Momentum*. We report *t*-statistics in parentheses. We double-cluster our standard errors at the analyst- and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	Earnings Forecast Change		Target Price Change		Recommendation Change	
	(1)	(2)	(3)	(4)	(5)	(6)
I(Alternative Data) × Δ	7.333***	7.620***	3.245***	2.567**	0.669**	0.600**
	(3.02)	(3.26)	(2.79)	(2.51)	(2.21)	(2.13)
Δ	3.785***	4.231***	2.286***	2.899***	0.712***	0.716***
	(3.63)	(4.58)	(5.50)	(6.61)	(8.95)	(8.91)
I(Alternative Data)	0.089*	0.105***	0.079*	0.071*	0.097**	0.104**
	(1.91)	(2.78)	(1.73)	(1.79)	(2.15)	(2.47)
Forecast Age	-0.029	-0.016	-0.028	-0.020	-0.028	-0.016
	(-1.49)	(-0.76)	(-1.48)	(-0.99)	(-1.52)	(-0.81)
Analyst/Firm Experience	0.001	-0.021*	0.001	-0.023**	0.001	-0.021*
	(0.44)	(-2.43)	(0.43)	(-3.52)	(0.67)	(-1.75)
Analyst Experience	0.001	0.017	0.001	-0.002	0.001	0.025
	(0.42)	(0.64)	(0.81)	(-0.08)	(0.71)	(1.21)
#Firms Covered	-0.055*	-0.092	-0.052	-0.071	-0.039	-0.085
	(-1.27)	(-1.39)	(-1.00)	(-1.07)	(-0.86)	(-1.37)
Forecast Frequency	0.018	0.090*	0.022	0.050	0.010	0.090**
	(0.74)	(2.47)	(0.78)	(1.43)	(0.43)	(2.40)
Broker Size	0.000	-0.001*	0.000*	-0.001*	0.000*	-0.001**
	(1.63)	(-1.68)	(1.75)	(-1.98)	(1.94)	(-2.13)
Firm Characteristics Controls	Yes	No	Yes	No	Yes	No
Analyst-Firm Fixed Effects	No	Yes	No	Yes	No	Yes
Firm-Year Fixed Effects	No	Yes	No	Yes	No	Yes
N	37,955	37,955	34,697	34,697	37,848	37,848
Adjusted R^2	0.023	0.045	0.024	0.046	0.024	0.044