THE USE AND USEFULNESS OF BIG DATA IN FINANCE: EVIDENCE FROM FINANCIAL ANALYSTS

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

This Draft: December 2023

Recent technological advances and the proliferation of alternative data vendors have created new information channels for investors. This paper investigates how these developments are reshaping the role of sell-side analysts. Employing textual analysis of analysts' written reports, we show that analysts themselves have begun to adopt alternative data into their analyses. Furthermore, we find that when analysts report drawing from alternative data, they generate more accurate earnings forecasts, and their brokerages subsequently receive higher amounts in trading commissions from investors, suggesting that investors value analysts' alternative data adoptions.

JEL Classification: G12, G14, G17, G24.

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

* Chi is affiliated with the Cornell SC Johnson College of Business, Cornell University. Hwang is affiliated with the Nanyang Business School, NTU, Singapore. Zheng is affiliated with the Alberta School of Business, University of Alberta. E-mail: fc354@cornell.edu, bh.hwang@ntu.edu.sg, and yaping2@ualberta.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

Sell-side analysts play a key role in financial markets. Their primary responsibility involves collecting and analyzing information on publicly traded companies and disseminating their insights to investors. To the degree that analysts' insights are unique and value-relevant, they can contribute significantly to enhancing market efficiency.

Like many roles in knowledge-driven sectors, the profession of sell-side analysts faces challenges. Part of this pressure is unique to the analyst field, stemming from a rise in passive investing and regulatory changes (Bradley, 2023). Another challenge emerges from a broader trend affecting a wide range of professions: the advancement of information technology and the corresponding surge in the availability of "big data" or "alternative data." 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 growing body of literature suggests that such data can predict companies' revenues and earnings and, thereby, help investors evaluate companies. Examples of alternative data that appear useful in predicting companies' performances include consumer activity estimated in real-time from mobile phone data (Froot, Kang, Ozik, and Sadka, 2017), online activity (Huang, 2018; Zhu, 2019) and satellite images of retail store parking lots (Zhu, 2019; Kang, Stice-Lawrence and Wong, 2021; Katona, Painter, Patatoukas and Zeng, 2023; Gerken and Painter, 2023).

The growing availability of such data, offering insights in a potentially more timely, comprehensive, and accurate manner than those traditionally provided by analysts, suggests a diminishing relevance of analysts to the investor community.

An alternative and perhaps initially counterintuitive perspective is that alternative data could reinstate analysts' role as pivotal information intermediaries and help them mitigate the negative impacts from passive investing and regulatory shifts. Suppose alternative data can occasionally provide unique and value-relevant insights above and beyond those available through traditional information sources. Suppose further that investors yearn to incorporate alternative data insights into their decision-making. One challenge that may prevent investors from doing so is the substantial cost associated with

subscribing to alternative data.[1] Relatedly, extracting meaningful insights from alternative data requires skill and experience.

Instead of each individual investor bearing the cost of subscribing to and learning about alternative data, it is more economical for a few analysts to incur these expenses and subsequently share their insights with investors. Analysts – who have the resources, readiness, and skill to study alternative data – could, therefore, utilize alternative data to reinstate their usefulness to investors rather than being made obsolete by alternative data.

Our paper examines this assertion. Our empirical analysis builds on the following idea: Analysts routinely publish written reports detailing their perspectives, along with the data and analyses underpinning these views. We posit that if analysts accessed alternative data and if the consideration of alternative data meaningfully altered their beliefs, they would discuss such influence in the corresponding reports. By parsing analysts' written reports and checking whether they explicitly reference the use of alternative data, we can thus gauge how often they draw from alternative data and in what context. We can also study investors' reactions to the issuance of such reports.

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 discuss the use of 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 is indeed couched in alternative data.

Due to the labor-intensive nature of our identification process, we limit our primary analysis to companies listed in the Dow Jones Industrial Average index (DJI) from June 2009 to May 2019. The

[1] E.g., https://www.institutionalinvestor.com/article/2bsx5x6m2pew8jzgavmyo/corner-office/how-much-are-managers-paying-for-data

DJI includes 30 large publicly traded companies. As the composition of the DJI changes over time, our final sample comprises 35 firms. We search for analyst reports on these 35 firms within the Investext database and integrate annual earnings forecast data from IBES. Ultimately, our dataset encompasses 64,018 written reports and their corresponding annual earnings forecasts, issued by 1,002 distinct analysts employed by 55 brokerage firms.[2]

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.

To gauge whether analysts are able to distill unique and value-relevant insights from alternative data, we test whether the annual earnings forecasts form reports couched in alternative data are more accurate than those from reports without alternative data references. We observe within a difference-in-differences specification that the incorporation of alternative data and annual earnings forecast accuracy are positively correlated with each other. Our estimates suggest that the performance improvement accompanying the consideration of alternative data is equivalent to having covered the corresponding firm for 3.6 additional years.

The foundation of the analyst business model rests on soft dollar arrangements, in which research services are funded through trading commissions. As investors perceive greater value in an analyst's research, they increase the number of trades routed through the analyst's brokerage and, as a result, pay a greater amount of commissions to the corresponding brokerage (Bradley, 2023). The perceived value of an analyst's research and the amount of trading commissions are thus directly linked with each other. To ascertain if analysts' adoption of alternative data into their analyses enhances their value to investors, we therefore examine whether discussions of alternative data positively correlate with the amount of commissions received.

[2] As we discuss in the main body of the paper, in additional analyses, we draw a random sample of 200 companies from the lower half of the size distribution within the CRSP database. We successfully retrieve 13,123 analyst reports for 143 out of the 200 firms, spanning from 2009 to 2019. We then examine to what degree our main findings, observed among the firms with the most extensive analyst coverage and the most active institutional investor involvement, extend to smaller firms.

To implement our test, we obtain institutional investor trade data from ANcerno, a firm that provides transaction cost analysis to institutional clients. The ANcerno dataset contains specific information for each institutional trade, including the commissions paid to the broker. For each written report, we consider all trades on the corresponding stock within three months after the report was issued and aggregate the total dollar value of the commissions paid to the corresponding broker. We observe within a difference-in-differences specification that institutional investors pay more in trading commissions when an analyst references alternative data. Our finding aligns with our supposition that analysts can enhance their value to investors by adopting alternative data.

Our final test examines the ramifications for investors and financial markets. We hypothesize that barriers to adopting alternative data are more pertinent for smaller or traditional institutional investors, such as mutual funds, than for hedge funds. Consistent with this conjecture, industry reports document that hedge funds are the primary consumers of alternative data.[3] By distilling insights from alternative data and sharing them with non-hedge-fund investors, analysts could play a crucial role in leveling the playing field between hedge funds and non-hedge-fund investors.

To assess this proposition, we again utilize our institutional investors' trade data. Existing research indicates that hedge funds place more profitable trades than other institutional investors (Stulz, 2007). Our paper examines whether this performance gap narrows when analysts incorporate alternative data and share their insights with investors, including non-hedge-fund investors.

To assess the performance of institutional investors, we construct transaction-based calendar time portfolios following the approach of Seasholes and Zhu (2010) and Ben-David, Birru, and Rossi (2019). As we are interested in trades affected by the dissemination of alternative data insights, we focus on transactions occurring within three months after the issuance of an analyst report and involve the stocks mentioned in these reports. We employ the methodology outlined by Jame (2018) to distinguish trades by hedge funds from those by other institutional investors.

[3] E.g., https://www.institutionalinvestor.com/article/2bsxas8ahvzgesgpnq0hs/portfolio/its-time-to-cash-in-on-big-data

First, we find that the stocks that hedge funds buy substantially outperform those they sell over the ensuing three months; the outperformance is 10.79% annualized. The 10.79% serves as our benchmark against which we compare the performance of non-hedge-fund institutional investors.

Our initial comparison considers cases where analysts do *not* reference alternative data. In these cases, we find that the stocks that non-hedge funds buy outperform those they sell by 3.85%.

The results are noticeably different when analysts reference alternative data. In such cases, non-hedge funds achieve a performance of 9.85%. The similarity in performance between hedge funds and non-hedge funds when analysts share their alternative data insights is consistent with the notion that analysts' use of alternative data helps level the playing ground between hedge funds and non-hedge-fund institutional investors.

Our paper contributes to several research areas. First, our paper relates to the literature on sell-side analysts. Analysts historically served as vital information intermediaries in financial markets and they have been the subject of extensive research. Bradley (2023) observes that the term "analyst" and its variations have appeared in the titles and abstracts of nearly 2,500 papers in the top finance and accounting journals over the last two decades.

An emerging body of work indicates that the analyst profession – already facing challenges from regulatory shifts and the growth of passive investing – is further threatened by technological advancements. There has been a surge in alternative data vendors (Grennan and Michaely, 2021). These data appear to predict companies' performances above and beyond those forecasted by human analysts (e.g., Froot, Kang, Ozik, and Sadka, 2017; Huang, 2018). Relatedly, Jame, Johnston, Markov, and Wolfe (2016) show that online platforms, which crowdsource investors' earnings forecasts and make the consensus forecasts publicly available, produce predictions that are more accurate than those provided by human analysts. Coleman, Merkley, and Pacelli (2022) study the emergence of "Robo-Analysts," which are "*research firms that focus on the use of technology to mass-produce recommendations with limited human involvement*" (page 12). The authors find that the recommendations of Robo-Analysts are superior to those of human analysts. Cao, Jiang, Wang, and Yang (2023) develop an "AI Analyst" by leveraging a range of machine-learning toolkits. Similar to Coleman et al. (2022), the authors find that their AI Analyst outperforms human analysts.

Collectively, the above studies point to a paradigm shift in the way investors can obtain information, with a trend toward advanced data analysis and big data. Our study adds to the analyst literature by providing evidence that some analysts respond to this trend by adopting alternative data themselves. Our results suggest that – by doing so – these analysts have maintained (or enhanced) their relevance to investors.

In that vein, our findings shed light on the broader question of how recent technological advances impact the labor market, particularly in knowledge-intensive industries (e.g., Agarwal, Gans, and Goldfarb, 2019; Frank et al., 2019). Our observations indicate a continued need for human labor, provided that individuals adapt to the evolving technological landscape. One possible reason is that humans can still occasionally distill insights that, as of now, are lost to machines. As a result, there is value in combining human and machine capabilities, a proposition that aligns with the findings of Cao, Jiang, Wang, and Yang (2023). Cao et al. (2023) find that while their AI Analyst outperforms human analysts, the most accurate predictions are theoretically achievable by integrating inputs from both AI and human analysts. Another possible and complementary explanation for the sustained need for human labor is the presence of barriers, which prevent many humans from adopting the technology. As long as these barriers exist, there will be opportunities for certain humans to specialize in these technologies and share their insights with the population who have not yet adopted them.

Finally, our study contributes to the body of research examining the consequences of alternative data for financial markets. The existing literature primarily studies the predictive value of alternative data for companies' future performances.[4] There is comparatively little research investigating the actual usage of alternative data by financial market participants. Current studies in this area typically take the initial offering of an alternative dataset as a positive shock to alternative data availability and compare financial market outcome variables from before to after the shock (Zhu, 2019). Other studies estimate financial market participants' use of alternative data for a specific stock by measuring how heavily the stock is covered in an alternative dataset (Dessaint, Foucault, and Fresard, 2023). In contrast, our

[4] Additionally, some studies suggest that econometricians can use alternative data, such as satellite images of retail store parking lots, to create proxies for local store performances (Kang, Stice-Lawrence, and Wong, 2021; Gerken and Painter, 2023). Researchers can then use these proxies to study to what degree local analysts or local investors rely on, or perhaps overemphasize, local information.

research considers situations where analysts explicitly describe their use of alternative data. One limitation of our method is that it might overlook situations where analysts consider alternative data but do not mention their use in their reports. The advantage of our approach is that we can document precisely what kinds of alternative data analysts report to use, for what firms, and in what situations. The descriptive evidence we present may serve as a valuable reference for future work on this topic.

2. Analysts' Use of Alternative Data

We begin the main body of our paper with a definition of alternative data. We then describe how we capture analyst reports that reference the use of alternative data.

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 large 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 (e.g., changes in household income); (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., weather data).

2.2 Measuring the Reliance on Alternative Data

To gauge analysts' use of alternative data, we conduct textual analysis on their written reports. We use the Investext database to download analyst reports for DJI constituents from June 1, 2009, through May 31, 2019. The Investext database contains active and historical research reports from brokerages, investment banks, and independent research firms around the globe. DJI constituents represent 30 large publicly traded firms. Because the DJI constituent list varies over time, our final sample comprises 35 firms.[5] For each report, we extract the ticker symbol, company name, report date, analyst names, broker name, report title, and 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 64,018 reports and earnings forecasts issued by 1,002 analysts from 55 brokers.

Just as some brokerages are missing in the IBES dataset (e.g., UBS), others are not included in the Investext database (e.g., Goldman Sachs). The non-comprehensive brokerage coverage raises questions regarding the generalizability of our findings to the broader population. Although we cannot

[5] 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 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.

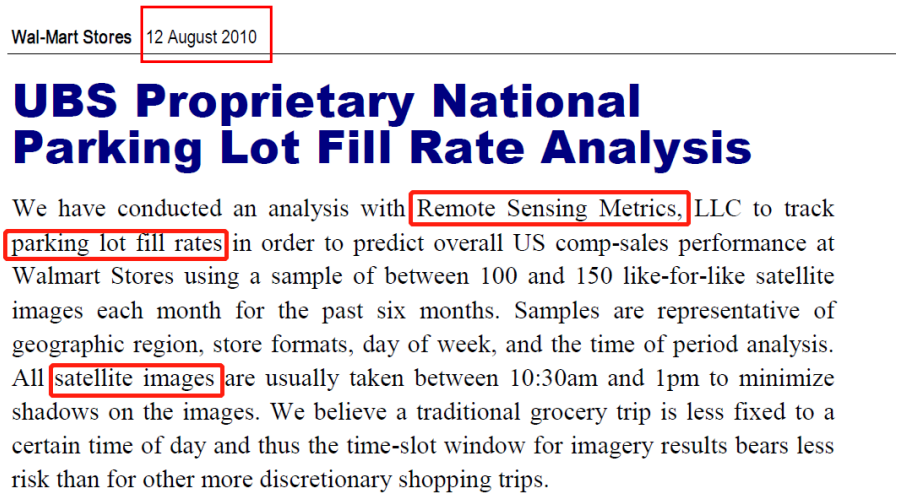
dismiss the potential for non-representative sampling, when we consider analyst reports and earnings forecasts on DJI constituents in the Investext and the IBES databases, respectively, we find no notable differences in underlying brokerage characteristics. The similarities in brokerage characteristics are evident in several key metrics: the number of analysts employed (81.26 vs. 71.07), the average number of firms each analyst covers (10.07 vs. 8.34), the average number of forecasts issued per month (4.84 vs. 4.15), and the average forecast accuracy (-0.54 vs. -0.52). At least based on these observables, there does not seem to be a systematic selection bias.

We 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.[6] To facilitate research on 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.[7] For each report identified by these initial keywords, we read the passages surrounding the initial keywords to verify that the report indeed adopts 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 reference an in-house data science team or external alternative data vendor and use alternative data, 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 incorporate 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 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 discuss the use of alternative data in their analyses.[8]

[6] In-house data science teams specialize in collecting and analyzing large unstructured data, which analysts can use in their valuation efforts.
[7] We convert all names and all text in the reports to lowercase characters.
[8] 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 adopt alternative data from those that do not.

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:[9]



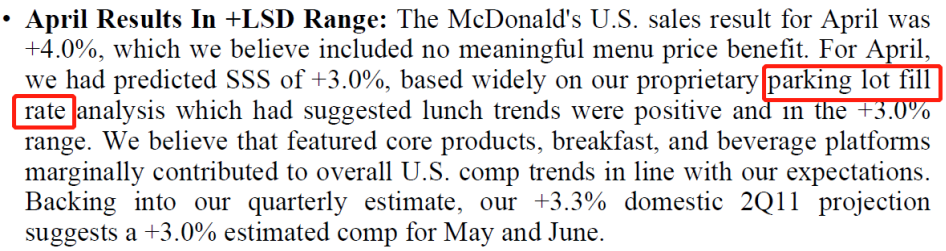
Wal-Mart Stores 12 August 2010

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 adopt alternative data but do not reference "Remote Sensing Metrics" or "RS Metrics."

The figure below is an excerpt from one such report:[10]



- **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 discussing the use of alternative data to verify that the analysts indeed incorporate alternative data into their reports. For instance, some firms in our sample provide satellite-related products or employ satellite imagery in their business

[9] UBS; Neil Currie, Krista Zuber, and David Eads; Walmart Inc; August 12, 2010.
[10] Piper Jaffary; Nicole Miller Regan and Joshua C. Long; McDonald's Corporation; May 9, 2011.

processes (e.g., oil and gas exploration). We exclude such cases. The figure below is an example of a false positive:[11]

> **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.

Our key variable turns to one if the following three conditions are met: (a) Analysts have access to alternative data. (b) Analysts believe the alternative data contain clear signals regarding a company's fundamentals, and they incorporate those signals into their forecasts and recommendations. (c) Finally, analysts disclose their reliance in their written reports. In Subsection 4.1, we discuss the implications of measuring this joint effect for the interpretation of our results.

Our variable may also turn to one if an analyst pretends to have studied alternative data. While we cannot rule out this possibility, such fabricated use would expose the analyst to serious career and litigation risks. Online Appendix Figure A2 provides an example of an analyst report that draws from alternative data. As depicted in the figure, discussions of alternative data often include visuals depicting its practical applications, such as satellite images of parking lots and their relation to revenues. After excluding standard boilerplate illustrations, the average number of figures included in analysts' reports is 0.86. When analysts make reference to the use of alternative data, this number rises to 4.16. The high level of detail typically provided by analysts when discussing their use of alternative data suggests further that it is improbable that analysts falsify their use of such data.

[11] Jefferies Group. John DiFucci, Joseph Gallo, and Howard Ma; Microsoft Corporation; May 11, 2017.

2.3 Descriptive Evidence Regarding Analysts' Discussion of Alternative Data

Our first test examines how frequently analysts report using 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 couched in reports that discuss the use of alternative data. By 2018/2019, the corresponding number is 10%.

The fraction of analysts discussing the use of alternative data for at least one of the firms they cover is naturally greater than the fraction of reports discussing alternative data use. In particular, we find that as of 2009/2010, 11% of the analysts in our sample discuss their use of alternative data. This fraction increases to 28% by 2018/2019.

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

In our study, we manually assign reports couched in 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 reference 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 couched in reports that discuss the use of 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%).[12]

Figure 1 displays two timelines. The first timeline indicates when – for our sample – 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 such 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, analysts extensively discuss the use of alternative data from six of the eight categories. Only app usage- and employee data are not widely adopted until mid-2015.

Overall, our evidence shows that analysts frequently adopt alternative data into their reports, at least for the largest and economically most meaningful firms.

3. The Usefulness of Alternative Data

In this section, we examine whether institutional investors, the primary clientele of analysts, value analysts' use of alternative data. We also consider the wider implications of analysts' alternative data adoptions on the quality of institutional investors' decision-making and the competitive balance among investors.

[12] Online Appendix Figure A3 shows how frequently alternative data from a particular category are discussed across industry sectors. Among others, the figure reveals that web traffic data are particularly frequently discussed for firms in Information Technology, whereas point-of-sale data are commonly references for firms in Consumer Discretionary.

3.1 Analysts' Use of Alternative Data and Forecast Accuracy

The hypothesis that institutional investors appreciate analysts' use of alternative data rests on the premise that analysts can derive unique and value-relevant insights from such data. To examine the validity of this assumption, we test whether earnings forecasts from reports couched in alternative data are more precise:

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

The observations are at the analyst/firm/forecast date level. 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 issued for firm *f* at time *t* and the corresponding 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* as of the corresponding forecast period, excluding analyst *i* and other analysts, who also report drawing from alternative data in their coverage of firm *f* as of time *t*. $PMAFE_{i,f,t}$ thus measures analyst *i*'s forecast accuracy relative to the forecast accuracies of all analysts, who cover the same firm at the same time, but do not report drawing from alternative data. 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 $PMAFE_{i,f,t} \times (-1)$.[13] We provide descriptive statistics regarding *Acc* and all other variables we use in this paper in Online Appendix Table A3.

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 reference alternative data.

[13] In robustness checks, we base our analysis on $AFE_{i,f,t}$ directly (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 $Acc_{i,f,t}$.

Controls include the following analyst characteristics: *Forecast Age*, *Analyst/Firm Experience*, *Analyst Experience*, *#Firms Covered*, *Forecast Frequency*, and *Broker Size*.[14] We do not control for firm characteristics as our fixed effects subsume them. Since our final sample comprises 64,018 written reports and earnings forecasts, the number of observations on which we estimate regression equation (1) is 64,018. We double-cluster our standard errors at the analyst- and year-month levels.

Difference-in-differences specifications often produce causal evidence. Our setting does *not* lend itself to causal inferences. That is, while the estimate of *I(Alternative Data)* indicates how much more accurate an analyst becomes in the post-adoption period relative to the pre-adoption period compared with her non-adopting counterparts, it cannot tell us how much of the abnormal performance improvement is truly caused by alternative data. The reason is that alternative data adoption is endogenous. For instance, 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").

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 explore whether the adoption of alternative data comes with greater effort. As detailed and tabulated in Online Appendix Table A4, alternative data adoption correlates neither with the timeliness of forecasts nor the number of forecast revisions. The adoption of alternative data also does not associate with the number of questions asked during earnings conference calls, the number of words spoken, or the types of questions asked; it marginally positively correlates with the number of conference calls attended.

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.

[14] We detail the construction of these variables in Appendix 2.

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 = 6.30). 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 0.060 (*t*-statistic = 2.72). Comparing the estimate of *I(Alternative Data)* with that of *Analyst/Firm Experience* suggests that the performance improvement accompanying the adoption of alternative data is equivalent to having covered the corresponding firm for 3.6 additional years.

Overall, while we cannot provide strong causal evidence, our results are at least consistent with the premise that analysts can derive distinct and value-relevant insights from alternative data.

3.2 Analysts' Use of Alternative Data and Trading Commissions

The analyst business model is rooted in soft dollar agreements. When institutional investors find more value in an analyst's research, they allocate a higher volume of trades through the brokerage that employs the analyst and, consequently, pay a greater amount in trading commissions to the corresponding brokerage. A simple test to determine if institutional investors place value on analysts' adoption of alternative data is thus to correlate the reported use of alternative data with the amount of trading commissions received.

We obtain institutional investor trade data from ANcerno, a firm that provides transaction cost analysis to institutional clients, including investment managers, like Fidelity Investments, and plan sponsors, like CalPERS (Hu, Jo, Wang, and Xie, 2018). While the ANcerno dataset covers a wide set of funds, the coverage is not comprehensive. Hu et al. (2018) estimate that the institutions in the ANcerno dataset account for 15% of all institutional trading volume. As discussed by Hu et al. (2018) and other studies utilizing the ANcerno dataset (Puckett and Yan, 2011; Jame, 2018), the dataset does not seem to suffer from sample selection bias.

For each institution covered by ANcerno, the dataset contains specific information for each of their trades, such as ticker symbol, date, trade direction, number of shares traded, and – crucially – identifiers for the investment manager, the broker executing the trade, and the commissions paid to the broker.

The ANcerno data contain records on all 35 DJI constituents. While the ANcerno data span the period from 1997 through 2015, starting from 2011, ANcerno no longer reports the identifier, which allows us to distinguish trades by individual institutions. Given that our sample of earnings forecasts and analyst reports commences in 2009, our brokerage commissions analysis thus encompasses the years 2009 and 2010. This period includes a total of 7,634 analyst reports. Of these, 2,877 are issued by brokerages that are not in the ANcerno database. Our sample for this part of our paper thus comprises 4,757 analyst reports.

In this study, we aim to test how the distribution of commissions is influenced by the dissemination of alternative data insights. Therefore, we consider all transactions within the first three months following a report's issuance that involve the stock covered in the report. We then total the commissions earned by the brokerage issuing the report. The median commissions earned by a brokerage in the first three months following a report's publication is $11,578.40. To the degree that ANcerno covers 15% of all institutional trades and the coverage is representative, we may estimate that the total median commissions earned by a brokerage in the first three months following a report's publication is $77,189.33. In additional analyses not shown here, we experiment with time frames other than a three-month period. The results from these tests are in line with those reported in this study and are available upon request.

We re-estimate regression equation (1) but replace the earnings forecast accuracy variable with our new trading commissions variable. The estimated coefficient β now indicates how much more trading commissions the analyst's brokerage receives in the post-adoption period relative to the pre-adoption period compared with analysts covering the same firm over the same period but not adopting alternative data. We again double-cluster our standard errors at the broker- and year-month levels.

As reported in Table 4, our estimate suggests that, on average, an alternative data report increases the total commissions paid to the brokerage by an additional $11,858.82 (*t*-statistic = 1.86), essentially doubling the median commissions earned by a brokerage.

Overall, our findings suggest that institutional investors recognize and value analysts' incorporation of alternative data and compensate them accordingly.

3.3 Analysts' Alternative Data Use and the Playing Field Among Institutional Investors

The adoption of alternative data by sell-side analysts has the potential to reshape not only the relationship between analysts and investors but also the competitive balance among the investors they serve. As alluded to in the introduction, hedge funds have been at the forefront of utilizing alternative data, while other institutional players, such as mutual funds, exhibit slower adoption rates. Additionally, existing research suggests that hedge funds execute more profitable trades than non-hedge funds. This strong performance may partially arise because hedge funds have superior data.

In this part of the study, we investigate whether the dissemination of alternative data insights through analyst reports has helped democratize access to these insights, thus leveling the playing field between hedge funds and non-hedge-fund investors. To assess this possibility, we again combine our analyst data with the ANcerno data and explore whether analysts' adoption of alternative data has lowered the performance gap usually seen between hedge funds and non-hedge funds.

As we are interested in trades affected by the dissemination of alternative data insights, we again focus on transactions that occur within three months after the issuance of an analyst report and that involve the stock discussed in the report. Analyses based on alternate windows produce results that are similar to the ones presented here and are available upon request. Applying Jame's (2018) methodology, we separate transactions by hedge funds from those made by other institutional investors.[15]

We further separate non-hedge-fund trades by whether they occur after an analyst discusses alternative data or whether they are placed without such accompanying reports. To ensure that the

[15] For more details, we refer the reader to Jame (2018).

investor composition behind the trades associated with alternative data and those without such linkage is comparable, we restrict our analysis to investors who appear in both subsets of trades.

We assess the performance of investors' trades by creating transaction-based calendar time portfolios, following the approach of Seasholes and Zhu (2010) and Ben-David, Birru, and Rossi (2019). Stocks purchased (sold) on day t are added to the buy (sell) portfolio at the beginning of day $t+2$. We assume a holding period of three months. We compute DGTW-adjusted returns for each stock/day and construct value-weighted portfolio returns. DGTW-adjusted returns are returns on a stock minus the value-weighted returns on a portfolio of stocks in the same size, book-to-market, and momentum quintile. We report the time-series average of the buy-minus-sell portfolio returns, whereby each daily portfolio return is weighted by the number of trades contributing to the portfolio on that specific day.

Table 5 reports the results. We find that the stocks hedge funds buy outperform those they sell. The difference in the stocks' ensuing returns is 10.79% annualized. The statistical significance is modest (t-statistic = 1.54), which is largely due to the high standard errors tied to the comparatively low number of hedge funds and the short sample period.

The performance is noticeably different for non-hedge funds. When analysts do not discuss the use of alternative data, the stocks that non-hedge funds buy outperform those they sell by 3.85% annualized. The annualized performance of 3.85% by non-hedge funds aligns with the findings of Puckett and Yan (2011) and Busse, Tong, Tong, and Zhang (2019), who also analyze ANcerno data.

Interestingly, when analysts incorporate alternative data, the performance of non-hedge funds improves significantly, reaching 9.85% annualized. This figure closely matches the performance achieved by hedge funds. Again, the standard error of the mean is very high due to the comparatively low number of analyst reports discussing the use of alternative data and the short sample period. Still, our results are at least consistent with the proposition that the employment of alternative data by analysts can equalize the competitive environment and allow the non-hedge funds that pay attention to alternative data insights to match the performance typically associated with hedge funds.

Another investor group that may benefit from analysts' alternative data adoption is retail investors. Although not the primary audience for analysts, retail investors can also access analyst reports, depending on their retail broker account and investment-related platforms subscribed.

To study the possible benefits retail investors might derive from analysts' use of alternative data, we utilize Trade and Quote (TAQ) data. We construct a measure of retail order imbalance following the method of Barber, Huang, Jorion, Odean, and Schwarz (2023). Our findings, presented in Online Appendix Table A5, reveal that retail investors are three times more likely to adjust their trades in line with changes in analyst recommendations when these recommendations come from reports incorporating alternative data. These results imply that at least some retail investors follow analyst recommendations anchored in alternative data more closely. To gauge the future performance of these trades, we estimate regressions of one-week, one-month, or three-month future returns on retail investor order imbalances. We distinguish between orders executed within the first two days following the release of an analyst report that discusses alternative data and those following the release of reports that do not include such discussions. As shown in Online Appendix Table A6, retail investor order imbalances more positively predict future returns when the orders are placed in the presence of alternative data-inclusive analyst reports. This finding implies that alternative data utilization assists retail investors in executing more profitable trades.

In separate analyses, we assess the broader market's reaction to analysts' use of alternative data. Specifically, we examine whether the market's response to changes in analyst research outputs, including earnings forecasts, price target forecasts, and overall recommendation levels, becomes more pronounced when those outputs are couched in alternative data. We detail our empirical design and our results in Online Appendix Table A7. In short, our results again suggest that broad sections of investors more closely follow analyst outputs when they are anchored in alternative data.

4. Discussion and Additional Analyses

Before concluding, we discuss why not more analysts explicitly reference the use of alternative data in their reports. We also examine the extent to which our findings regarding the largest U.S. firms apply

to smaller firms. Finally, we explore variations in our key independent variable, *I(Alternative Data)*, and our main empirical specification.

4.1 Why Don't Analysts More Frequently Discuss the Use of Alternative Data?

The seeming benefits accompanying the adoption of alternative data, such as increased trading commissions, raise the question of why analysts do not report the use of alternative data more frequently.

Our key independent variable, *I(Alternative Data)*, can be zero for at least three reasons. First, analysts may not have the resources or the willingness to study alternative data, causing *I(Alternative Data)* to be zero ("resource limitations"). Second, even if analysts access and study alternative data, they may not always uncover clear, unique, and value-relevant insights. Analysts are unlikely to discuss such "failed" use of alternative data in the limited space that is available to them in their reports ("intermittent usefulness"). Finally, analysts may identify relevant signals but strategically choose not to detail these insights in their reports ("strategic considerations").

4.1.1 The Relevance of Resource Limitations and Intermittent Usefulness

To gauge the relevance of the resource limitations and the intermittent usefulness perspectives in explaining why *I(Alternative Data)* does not equal one more frequently, we estimate the following probit regression:

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

The observations are at the analyst/firm/forecast date level. *Controls* include the following analyst and firm characteristics: *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.

Our first two key independent variables, *X*, are an indicator of whether an analyst's brokerage has an in-house data science team and the number of an analyst's colleagues from the same city already discussing the use of alternative data in one of their reports as of the analyst's forecast date. The expenses necessary to access and analyze alternative data are presumably reduced when analysts are

supported by an in-house data science team. Moreover, suppose individuals in knowledge-based industries owe much of their success to their colleagues (Hwang, Liberti, and Sturgess, 2019). In that case, analysts surrounded by peers already utilizing alternative data should find the learning curve less steep. If resource limitations are an important reason *I(Alternative Data)* does not equal one more frequently, we should observe more frequent alternative data adoptions when there are fewer resource limitations; we should thus observe positive coefficient estimates for the first two key independent variables.

Similarly, if intermittent usefulness is an important reason *I(Alternative Data)* does not equal one more frequently, we should observe a positive correlation between the frequency of alternative data adoptions and measures of its usefulness.

Alternative data have three key advantages. First, alternative data provide immediate indications of a company's performance, allowing analysts to draw from current rather than outdated information. Second, the origin of alternative data from independent third-party vendors mitigates the risk of bias or distortion often associated with data produced internally by company managers. Third, the detailed nature of alternative data, for instance, breaking down information to specific products or branches, provides analysts with a more detailed and nuanced understanding of a company's operational performance.

Building on these considerations, we propose that alternative data are more useful and, as a result, become more frequently discussed in analyst reports (1) when receiving instantaneous signals regarding a company's performance is critical, (2) when traditional data are ambiguous, and (3) when analysts lack access to granular data.

Receiving instantaneous signals regarding a company's performance becomes critical when there are relatively few company announcements and when the uncertainty regarding a company's performance is high. We thus conjecture that the usefulness of alternative data and the likelihood of alternative data adoption 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. For a description of the precise construction of these and the subsequent key independent variables outlined below, we again direct the reader's attention to Appendix 2.

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 traditional data, we consider whether the company has had to restate its earnings (Wilson, 2008) and the absolute value of discretionary accruals (Bhattacharya, Desai, and Venkataraman, 2013).

Finally, private meetings with management are one of the key channels through which analysts can obtain a more nuanced perspective of a company's performance (Green, Jame, Markov, and Subasi, 2014; Soltes, 2014; Solomon and Soltes, 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 an opportunity to interact with senior corporate managers, provide insight into whether a particular analyst has preferential access to the firms participating in the conference.

We report our results in Table 6. We find that analysts more frequently adopt alternative data into their reports when their brokerage has an in-house data science team and when analysts have colleagues already incorporating alternative data. These results are consistent with the resource limitations perspective.

Consistent with the intermittent usefulness idea, we find that analysts more frequently adopt alternative data when the company provides relatively few announcements, when stock return volatility and the absolute value of earnings surprises are high, when the firm has had to restate its earnings and the absolute value of discretionary accruals are high and when analysts lack preferential access to management through private meetings.[16]

[16] In Online Appendix Table A8, we describe analyses showing that the incremental improvements in earnings forecast accuracy associated with discussions of alternative data use also strengthen when the absolute value of earnings surprises are high and when the firm has had to restate its earnings.

Overall, our results suggest that resource limitations and intermittent usefulness are important determinants of *I(Alternative Data)* and help explain why analysts do not discuss alternative data in their reports more frequently.

4.1.2 The Relevance of Strategic Considerations

Analysts may choose not to disclose their reliance on alternative data for two strategic reasons. First, analysts may strive to continuously provide fresh perspectives. An analyst reporting the use of alternative data in one year may thus avoid mentioning her continued reliance in the following years to avoid repetition.

To investigate this possibility, we consider cases where analysts explicitly reference the use of alternative data for a particular firm in a particular year. We find that these analysts discuss their reliance on the same alternative data category in the subsequent year 55% of the time.

It appears likely that an analyst who incorporated alternative data into her report in one year could easily obtain an updated version of the data in the ensuing year. The fact that 45% of the times, analysts do not reference the same alternative data category in the subsequent year suggests either that the perceived value of alternative data decreases rapidly or that analysts avoid repeating their continued reliance.

To shed light on the relevance of these possibilities, we modify our earnings forecast accuracy regression. Our adjusted independent variable equals one if an analyst does not discuss the use of alternative data in year *t* but did so in year *t*-1. Suppose analysts no longer mention the use of alternative data because they find the data no longer useful. Suppose further that analysts' assessments of the diminished usefulness are accurate. In that case, our revised independent variable should no longer be significantly associated with earnings forecast accuracy.

Our analysis yields a coefficient estimate of 0.070 (*t*-statistic = 2.57). This robust estimate is consistent with the idea that analysts still benefit from alternative data insights but opt not to reiterate their reliance to prevent redundancy. At the same time, the estimate of 0.070 is markedly lower than that of our main specification (coefficient estimate = 0.214, *t*-statistic = 6.30), suggesting that, in many cases, the utility of alternative data does diminish over time.

Another reason analysts may not disclose their use of alternative data for strategic considerations is that analysts fear the disclosure of comprehensive methodologies and data sources makes it easier for competing analysts to imitate and employ analogous analytical techniques. This could result in the erosion of the unique competitive advantage initially held by the original analyst.

In our sample, we detect only 92 cases where the initial adoption of alternative data is replicated by another analyst in the concurrent year for the identical company. In these situations, we find that the original analyst, on average, receives $20,059.55 in commissions through trades in the corresponding stock occurring within the first three months of the alternative data report publication. Subsequent to the imitation, the total three-month average commission of the original analyst stands at $25,537.25, suggesting that the financial repercussions from imitation are limited.[17] Overall, our results suggest that strategic considerations play some, but not a critical role in explaining why *I(Alternative Data)* does not equal one more frequently.

4.2 The Use and Usefulness of Alternative Data Among Small Firms

Our tests so far have been on constituents of the DJI. To explore the prevalence and impact of alternative data within the small-cap sector, we analyze a random sample of 200 companies drawn from the lower half of the size distribution in the CRSP database, covering the years from 2009 to 2019. Out of these companies, 143 have earnings forecasts and analyst reports data. The median market capitalization of our small firms is $1.013 billion. In comparison, the median market capitalization of DJI constituents is $11.854 billion. We retrieve a total of 13,123 analyst reports for our small firms over the 2009-2019 period, compared to 64,018 reports for the 35 firms in the DJI. Our small firms thus receive substantially lower coverage, which should not surprise given that analysts' key clientele, institutional investors, primarily invest in larger, more liquid stocks.

Following the same procedure described in Subsection 2.2, we identify the fraction of reports discussing the use of alternative data. We find that for our small firms, analysts discuss alternative data in only 2% of their reports, compared with 9% for the DJI constituents. It thus appears that analysts use

[17] Concerns about imitation might still affect analysts' reporting behavior, if analysts are unaware that imitations are rare and do not affect an original analyst's trading commissions.

alternative data significantly less frequently among small firms. One possible explanation is that analysts are reluctant to bear the financial and learning costs associated with alternative data adoption for firms that attract limited attention from institutional investors. Another possible reason is that the quality of alternative data is lower for smaller firms.

We find that the reference of alternative data correlates with more accurate earnings forecasts even among our small firms. As reported in Online Appendix Table A9, the coefficient estimate is 0.197, suggesting that an analyst at the median accuracy level improves to the 60th percentile. The magnitude of the association is similar to that for DJI constituents, as our results for DJI constituents suggest that the adoption of alternative data elevates an analyst from the median to the 62nd percentile in terms of earnings forecast accuracy.

We also re-run our analyses on the trading commissions received by analysts' brokerages and the performance differential between hedge funds and non-hedge-fund institutional investors. Since institutions trade significantly less in smaller firms, the sample sizes for these additional analyses are substantially smaller. As shown in Online Appendix Table A10, we find evidence that institutional investors continue to reward analysts who discuss alternative data by directing more trades through their brokerages. The coefficient estimate is 2,818.23 (*t*-statistic = 3.38). Compared to the estimate of 11,858.82 for DJI constituents, this estimate is substantially smaller in magnitude. The weaker economic significance suggests that there is less incentive for analysts to adopt alternative data for smaller firms, which could explain why the fraction of reports couched in alternative data is so much lower for smaller firms than for DJI constituents.

When we repeat our transaction-based calendar time portfolio analysis for the smaller firms, we find that analysts' discussions of alternative data coincide with a reduced performance gap between institutional investors and hedge funds. However, the standard errors are so large that none of the average returns are statistically different from zero.

4.3 Differences in the Usefulness by Alternative Data Types

4.3.1 Variation Across Alternative Data Categories

While our results suggest that alternative data contain unique and value-relevant insights, the usefulness of the data may vary by its type. To explore this possibility, we first replace *I(Alternative Data)* with eight indicator variables, each denoting whether an analyst uses alternative data from a particular alternative data category. We then re-estimate our earnings forecast accuracy regression (1). The results in Table 7 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 (7) geospatial data and (8) 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.[18] 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.[19]

In another test, we replace *I(Alternative Data)* with the number of distinct alternative data categories discussed in an analyst's report. We then re-estimate our regressions within the subset of cases where an analyst reports the use of alternative data from at least one category. In short, we detect a strong positive correlation between the number of alternative data categories and the accuracy of earnings forecasts. This finding implies that when analysts' views are supported by a wider range of data types, their predictions are particularly accurate (Table 7).

[18] The survey results are viewable at https://alternativedata.org/stats/.

[19] In separate tests, we also gauge the usefulness of different alternative data categories across industries. Specifically, we re-estimate our regressions separately for each Global Industry Classification Standard Sector. We present the results in Online Appendix Figure A4. Among others, our results suggest that app-usage data is particularly benefial for forecasting the performance of firms in the information technology sector, while point-of-sale data is particularly advantageous for predicting the performance of firms in the consumer staples industry.

4.3.2 Variation Across More or Less Proprietary Alternative Data

Among our eight alternative data categories, four categories: (1) sentiment data, (2) employee data, (3) geospatial data, and (4) web traffic data, are comparatively more accessible to the public and less dependent on external data vendors ("accessible data"). For example, investors can acquire variants of sentiment data through social media platforms, catch a glimpse of employee-related information through public websites (e.g., Glassdoor.com), gather geospatial data through a combination of Google Maps and publicly available demographic data, and retrieve web traffic data using tools, such as Google Trends. In contrast, data from the remaining four categories: (5) app-usage data, (6) point-of-sale data, (7) satellite image data, and (8) other unspecified types of data, are generally not accessible to the general public ("proprietary data").[20]

In separate tests, we gauge whether accessible data are similarly useful to analysts in predicting a company's performance as proprietary data. Our dataset reveals that among the analyst reports that discuss the use of alternative data, 64.2% base their analysis on accessible data, while 49.3% rely on proprietary data.[21] Further examination of the relationship between analyst characteristics and the use of accessible- and proprietary data shows that analysts employed by larger brokerages or brokerages with internal data science teams, along with analysts whose colleagues discuss their use of alternative data, are more inclined to reference proprietary data.

Regarding the implications for forecast accuracy, Table 7 shows that references of proprietary data (coefficient estimate = 0.243, *t*-statistic = 6.82) and accessible data (coefficient estimate = 0.188, *t*-statistic = 4.07) come with similar increases in the precision of earnings forecasts; the two coefficient estimates are not statistically different from each other (*F*-statistic = 0.99, *p*-value = 0.32).

[20] We emphasize that this is a relative statement. Even data types we deem as more accessible to the public, like web-traffic data, encompass specific details such as the duration of user visits to a merchant's website, which are typically not accessible to the general public.

[21] The sum of the two percentages exceeds 100% as some analyst reports base their analysis on both accessible and proprietary data.

4.4 Alternative Data Use and Earnings Forecast Accuracy: Instrumental Variable- and Matching Analyses

To study the relationship between alternative data use and earnings forecast accuracy, our primary analysis adopts a standard difference-in-differences method and assesses 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 reference alternative data. As alluded to in Subsection 3.1, an analyst's decision to adopt alternative data is not exogenous and, instead, could reflect the decision to exert greater effort. To speak to this concern, we experiment with an instrumental variable approach. The basic premise behind our instrumental variable approach is that an individual analyst's ability to use alternative data is contingent upon her firm's infrastructure and technology investment ("relevance"). At the same time, certain brokerage level decisions, which lead to easier access to alternative data for analysts, are independent of a single analyst's preferences or actions ("exclusion restriction").

We consider two instruments. The first, *First Time Use*, is an indicator variable, which turns to one when an analyst's colleague, working in the same city, adopts alternative data for the first time. The second, *Software Budget*, refers to the allocated budget for software purchases at the broker-year level, sourced from Aberdeen's Computer Intelligence Technology Database. We assess the validity of these instruments through the Hansen *J*-test for overidentifying restrictions and confirm their efficacy with underidentification tests and *F*-tests in the first-stage regression. Our regression results pass all diagnostic tests, implying that our instrumenting strategy is valid.

We report the results from the instrumental variable approach in Online Appendix Table A11. Column (1) reports the results from the first-stage test. The coefficient estimates of *First Time Use* and *Software Budget* are positive and significant at the 1% level, suggesting that our instruments are highly correlated with the endogenous variable. Column (2) shows that the second-stage coefficient estimates on the instrumented *I(Alternative Data)* are positive and significant. The estimates are similar to those from the original regression specification.

We also experiment with a matching sample analysis. We pair each alternative data report within our sample with a non-alternative data report written by analysts covering the same firm during

the same forecast period. We ensure that the analysts of the matched reports work for brokerages in the same size quintile (based on the number of analysts employed) and the same firm-specific experience quintile (based on the number of years of having covered the respective firm). If multiple reports satisfy the above criteria, we select the report with the most similar forecast horizon as the alternative data report. We then repeat our forecast accuracy analysis within this matching sample.

The results of this matching sample approach are presented in Online Appendix Table A12. The coefficient estimate for *I(Alternative Data)* equals 0.204 with a *t*-statistic of 3.83. Again, this estimate is very similar to that from our primary analysis.

4.5 Other Key Performance Indicators

Our final test considers revenue forecast accuracy as another crucial Key Performance Indicator for analysts. Revenue forecasts are sourced directly from the IBES database, and the accuracy of these forecasts is measured in the same manner as earnings forecast accuracy. Since not all analysts have revenue forecasts in the IBES database, our sample size drops to 27,661.

In addition to revenue forecast accuracy, one might also consider the accuracy of cost forecasts. Unfortunately, the IBES database does not contain cost forecasts. Instead, we compute "residual forecasts" by taking the difference between revenue-per-share forecasts and earnings-per-share forecasts and constructing a measure of accuracy based on these residual forecasts. The resulting measure may be seen as capturing any improvement in earnings forecast accuracy that cannot be tied to more accurate revenue forecasts.

We report our results in Online Appendix Table A13. Column (1) presents results based on revenue forecast accuracy, and column (2) focuses on residual forecast accuracy. The results in column (1) reveal that discussing alternative data associates with significantly more accurate revenue forecasts, as evidenced by the significance and substantial magnitude of the coefficient estimate of *I(Alternative Data)*. Conversely, the findings in column (2) suggest that the use of alternative data does not lead to more accurate residual forecasts, given that the estimate of *I(Alternative Data)* is close to zero. Put differently, our results indicate that once any improvement tied to more accurate revenue forecasts is removed from the analysis, the adoption of alternative data no longer leads to more accurate earnings

forecasts. Such a suggestion seems reasonable, considering that most alternative data pertain to a company's sales, not profits.

5. Conclusion

Our study documents the dynamic role of sell-side analysts. Rapid advancements in information technology and the emergence of alternative data sources are creating a wealth of information, which, from an investor's point of view, might dominate those provided by sell-side analysts, eventually rendering the analyst profession obsolete.

Our paper points to a more nuanced picture. We propose that the complexity and costs associated with accessing and interpreting alternative data represent a significant challenge for many investors. This challenge opens a window of opportunity for analysts to serve as essential conduits of information. Instead of investors each bearing the financial and learning costs associated with studying alternative data, it is more efficient for a few analysts to absorb these expenses and become proficient in parsing alternative data. Analysts can then share their insights with investors, who, in turn, can integrate the signals with their other pieces of information to determine the value of a stock.

Our research supports this perspective. We find that analysts have begun to adopt alternative data and that investors value analysts' alternative-data-driven insights and reward them accordingly. Our study thus highlights the enduring importance of human expertise in financial analysis, even in an era increasingly reliant on big data and technology. Simultaneously, our findings indicate that maintaining this relevance requires a transformation in the function and approach of human analysts.

References

Agrawal A, Gans JS, Goldfarb A (2019) Artificial intelligence: the ambiguous labor market impact of automating prediction. *J. Econom. Perspectives*. 33(2): 31-50.

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

Barber BM. Huang X, Jorion P, Odean T, Schwarz C (2023) A (sub) penny for your thoughts: Tracking retail investor activity in TAQ. *J. Finance*. Forthcoming.

Ben-David I, Birru J, Rossi A (2019) Industry familiarity and trading: Evidence from the personal portfolios of industry insiders. *J. Financial Econom*. 132(1): 49-75.

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 (2023) Financial Analysts. *Handbook of Financial Decision Making*. (Edward Elgar Publishing, Northampton, MA)

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.

Busse JA, Tong L, Tong Q, Zhang Z (2019) Trading regularity and fund performance. *Rev. Financial Stud*. 32(1): 374-422.

Cao S, Jiang W, Wang JL, Yang B (2023) From man vs. machine to man + machine: The art and AI of stock analyses. Working paper, National Bureau of Economic Research, Cambridge, MA.

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

Coleman B, Merkley K, Pacelli J (2022) Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *Accounting Rev*. 97(5): 221-244.

Dessaint O, Foucault T, Frésard L (2023) Does Alternative Data Improve Financial Forecasting? The horizon effect. *J. Finance*. Forthcoming.

Frank MR, Autor D, Bessen JE, Brynjolfsson E, Cebrian M, Deming DJ, Feldman M, Groh M, Lobo J, Moro E, Wang D (2019) Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*. 116(14): 6531-9.

Froot K, Kang N, Ozik G, Sadka R (2017) What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *J. Financial Econom*. 125(1): 143-162.

Gerken WC, Painter MO (2023) The value of differing points of view: Evidence from financial analysts' geographic diversity. *Rev. Financial Stud*. 36(2): 409-449.

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.

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.

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.

Hu, G., Jo, K.M., Wang, Y.A. and Xie, J., (2018). Institutional trading and Abel Noser data. *Journal of Corporate Finance*, *52*, pp.143-167.

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

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

Jame R (2018) Liquidity provision and the cross section of hedge fund returns. *Management Sci*. 64(7): 3288-3312.

Jame R, Johnston R, Markov S, Wolfe MC (2016) The value of crowdsourced earnings forecasts. *J. Accounting Res*. 54(4): 1077-1110.

Kang JK, Stice-Lawrence L, Wong YTF (2021) The firm next door: Using satellite images to study local information advantage. *J. Accounting Res*. 59(2): 713-750.

Katona Z, Painter M, Patatoukas PN, Zeng J (2023) On the capital market consequences of alternative data: Evidence from outer space. *J. Financial Quant. Anal*. Forthcoming.

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.

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.

Puckett A, Yan X (2011) The interim trading skills of institutional investors. *J. Finance.* 66(2): 601-633.

Seasholes MS, Zhu N (2010). Individual investors and local bias. *J. Finance.* 65(5): 1987-2010.

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.

Stulz RM (2007) Hedge funds: Past, present, and future. *J. Econom Perspectives.* 21(2): 175-194.

Wigglesworth R (2016) Investors mine big data for cutting-edge strategies. *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.

Zhu C (2019) Big data as a governance mechanism. *Rev. Financial Stud*. 32(5): 2021-2061.

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 and to gauge inventory levels 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, weather 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. The average absolute forecast error is calculated across all analysts covering firm *f* as of the corresponding forecast period, excluding analyst *i* and other analysts, who also report drawing from alternative data in their coverage of firm f as of time *t*. 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.
Trading Commissions	For each written report, we consider all trades on the corresponding stock within three months after the report was issued and aggregate the total dollar value of the commissions paid to the corresponding broker.
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 in the same city as the corresponding analyst.
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.

Appendix 2. Continued.

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.
Size	The market capitalization of the corresponding firm at the end of the previous fiscal year in billions.
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.
I(Category = X)	An indicator variable that equals one if the analyst explicitly references the use of alternative data from alternative data category X.
$\sum$ *Categories*	The number of different alternative data categories an analyst references in her report.
I(Source = Proprietary Data)	An indicator variable that equals one if the analyst explicitly references the use of proprietary alternative data.
I(Source = Accessible Data)	An indicator variable that equals one if the analyst explicitly references the use of accessible alternative data.

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.

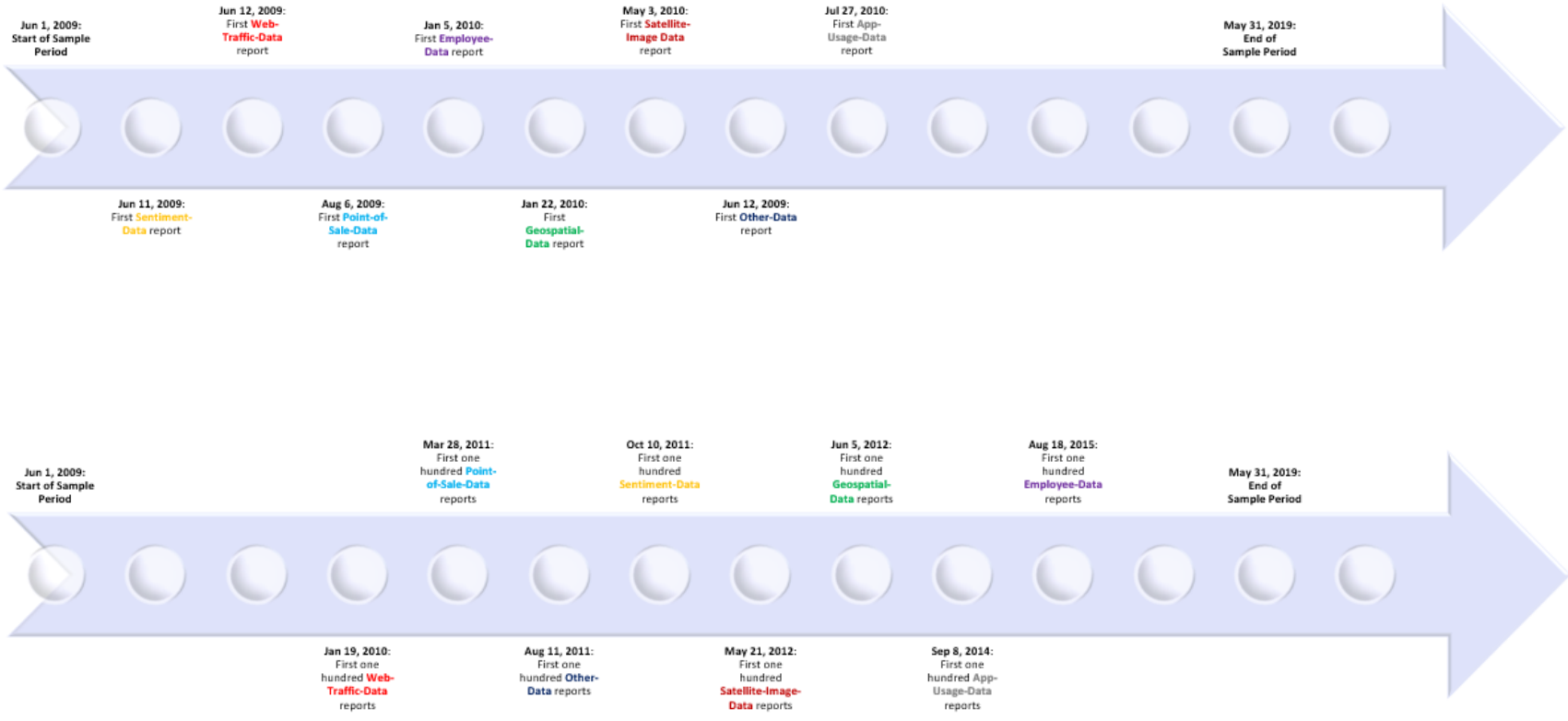


Table 1
Number and Fraction of Analyst Forecasts Explicitly Supported by Alternative Data

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	7,616	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	58,379	9%
Panel B: By Industry Sector			
Energy	40	2,512	2%
Materials	29	2,596	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 not referencing the use of alternative data, 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. 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)
I(Alternative Data)	0.214***
	(6.30)
Forecast Age	-0.251***
	(-12.34)
Analyst/Firm Experience	0.060***
	(2.72)
Analyst Experience	0.056
	(1.00)
#Firms Covered	0.029
	(0.52)
Forecast Frequency	0.033
	(1.11)
Broker Size	-0.000
	(-0.85)
Analyst-Firm Fixed Effects	Yes
Firm-Year Fixed Effects	Yes
N	64,018
Adjusted R^2	0.238

Table 4
Alternative Data and Trading Commissions

This table reports coefficient estimates from regressions of trading commissions paid to the brokerage on whether an analyst explicitly references the use of alternative data in her corresponding written report. The observations are at the brokerage/firm/forecast date level. We consider trades and the associated trading commissions in the ANcerno database. As we are interested in how the allocation of trades and commissions are affected by the dissemination of alternative data insights, we focus on transactions, which occur within three months after the issuance of an analyst report and which involve the stock discussed in the report. For each analyst report, we compute the total commissions earned by the brokerage issuing the report and use this as our dependent variable. *I(Alternative Data)* equals one if the corresponding reports explicitly describes the use of alternative data as described in Section 2.2 and zero otherwise. We define all remaining variables in Appendix 2**.** We report *t*-statistics in parentheses. We double-cluster our standard errors at the broker- and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	(1)
I(Alternative Data)	11,858.82*
	(1.86)
Forecast Age	-3,393.30
	(-1.15)
Analyst/Firm Experience	1,416.47
	(0.63)
Analyst Experience	1,233.77*
	(1.88)
#Firms Covered	5,091.09
	(0.22)
Forecast Frequency	-2,909.48
	(-0.33)
Broker Size	-720.16***
	(-4.23)
Broker-Firm Fixed Effects	Yes
Firm-Year Fixed Effects	Yes
N	4,757
Adjusted R^2	0.623

Table 5
Alternative Data and Portfolio Returns

This table reports the performance of transaction-based buy-and-sell calendar-time portfolios. We consider trades in the ANcerno database. As we are interested in trades affected by the dissemination of alternative data insights, we focus on transactions, which occur within three months after the issuance of an analyst report and which involve the stock discussed in the report. Applying Jame's (2018) methodology, we separate trades by hedge funds from those made by other institutional investors. We further separate non-hedge-fund trades by whether trades occur after the analyst discusses alternative data or whether trades are placed without such accompanying reports. We assess the performance of trades by creating transaction-based calendar time portfolios, following the approach of Seasholes and Zhu (2010) and Ben-David, Birru and Rossi (2019). Stocks purchased (sold) on day t are added to the buy (sell) portfolio at the beginning of day $t+2$. We assume a holding period of three months. We compute DGTW-adjusted returns for each stock/day (= return on a stock minus the value-weighted return on a portfolio of stocks in the same size, book-to-market, and momentum quintile) and construct value-weighted portfolio returns. We report the time-series average of the buy-minus-sell portfolio returns, whereby each daily portfolio return is weighted by the number of trades contributing to the portfolio on that specific day. We also report the annualized difference. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

	Buy – Sell [Daily] (1)	Buy – Sell [Annual] (2)
Hedge funds	0.04%	10.79%
	(1.54)	
Non-Hedge Funds without Alternative Data	0.02%**	3.85%
	(2.19)	
Non-Hedge Funds with Alternative Data	0.04%	9.85%
	(1.27)	

Table 6
Variation in the Use of Alternative Data

This table reports coefficient estimates from regressions of alternative data adoption on various analyst- and firm characteristics. The observations are at the analyst/firm/forecast date level. Column (1) reports the results from a probit model, whereas Column (2) reports the results from a linear probability model to allow for consistent parameter estimation while including fixed effects. The dependent variable equals one if the analyst's earnings forecast is explicitly supported by alternative data and zero otherwise. *I(In-House Data Science Team)* equals one if the analyst's brokerage has an in-house data science team. $\sum$ *Colleagues* $_{Alternative\ Data}$ is the number of colleagues working in the same city as the corresponding analyst. *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 $\sum$ *Colleagues* $_{Alternative\ Data}$, *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)
I(In-House Data Science Team)	0.514***	0.104***
	(4.42)	(3.55)
Rank($\sum$*Colleagues* $_{Alternative\ Data}$*)*	0.068**	0.008
	(2.15)	(1.48)
Rank(Number of 8-Ks)	-0.100***	-0.012**
	(-3.30)	(-2.20)
Rank(Return Volatility)	0.113***	0.011**
	(4.02)	(2.54)
Rank(Earnings Surprise)	0.057***	0.009***
	(2.76)	(3.04)
I(Earnings Restatement)	0.300***	0.072***
	(3.08)	(3.55)
Rank(Discretionary Accruals)	0.042*	0.010***
	(1.72)	(2.85)
I(Lack of Preferential Access to Management)	0.160	0.028*
	(1.61)	(1.93)
Analyst/Firm Experience	-0.006	-0.001
	(-0.81)	(-0.91)
Analyst Experience	0.004	0.001
	(0.60)	(0.98)
#Firms Covered	0.135	-0.005
	(1.00)	(-0.29)
Forecast Frequency	-0.024	0.005
	(-0.37)	(0.58)
Broker Size	0.001	0.000
	(0.99)	(1.47)

Table 6. Continued.

	(1)	(2)
Size	0.210***	0.046***
	(3.43)	(3.57)
M/B	0.012	0.001
	(1.29)	(0.54)
Momentum	0.175	0.027
	(1.25)	(1.24)
Analyst-Firm Fixed Effects	No	No
Firm-Year Fixed Effects	No	No
Industry Fixed Effects	No	Yes
N	64,018	64,018
Pseudo R^2	0.118	0.122

Table 7
Differences in the Usefulness by Alternative Data Types

This table reports results from analyses similar to those in Table 3. In Column (1), we replace *I(Alternative Data)* with *I(Category = X)*, which equals one if the analyst explicitly references the use of alternative data from alternative data category X. In Column (2), we replace *I(Alternative Data)* with ∑*Categories*, which is the number of different alternative data categories an analyst references in her report. In Column (3), we replace *I(Alternative Data)* with *I(Source = Proprietary Data)* and *I(Source = Accessible Data)*, which equal one if the analyst explicitly references the use of proprietary and accessible alternative data as described in Section 4.3.2, respectively. 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)	(3)	(4)
I(Category = App Usage)	0.310***			
	(4.04)			
I(Category = Sentiment)	0.230***			
	(2.88)			
I(Category = Employee)	0.212***			
	(3.09)			
I(Category = Geospatial)	-0.033			
	(-0.31)			
I(Category = Point of Sale)	0.183***			
	(3.44)			
I(Category = Satellite Image)	0.053			
	(0.53)			
I(Category = Web Traffic)	0.137**			
	(2.09)			
I(Category = Others)	0.187***			
	(3.88)			
∑ *Categories*		0.175*		
		(1.75)		
I(Source = Proprietary Data)			0.243***	
			(6.82)	
I(Source = Accessible Data)				0.188***
				(4.07)
Forecast Age	-0.250***	-0.181***	-0.251***	-0.252***
	(-12.36)	(-3.67)	(-12.19)	(-12.31)
Analyst/Firm Experience	0.060***	-0.017	0.059**	0.061***
	(2.68)	(-1.24)	(2.60)	(2.79)
Analyst Experience	0.055	0.378	0.059	0.057
	(0.98)	(1.50)	(1.03)	(1.02)
#Firms Covered	0.030	-0.295	0.029	0.034
	(0.54)	(-1.10)	(0.53)	(0.60)
Forecast Frequency	0.033	-0.101	0.033	0.029
	(1.09)	(-0.96)	(1.08)	(0.96)
Broker Size	-0.000	0.004	-0.000	-0.000
	(-0.91)	(1.46)	(-0.84)	(-0.86)

Table 7. Continued.

	(1)	(2)	(3)	(4)
Analyst-Firm Fixed Effects	Yes	Yes	Yes	Yes
Firm-Year Fixed Effects	Yes	Yes	Yes	Yes
N	64,018	5,639	64,018	64,018
Adjusted R^2	0.239	0.406	0.237	0.236