


On the Capital Market Consequences of Big Data: Evidence from Outer Space

Zsolt Katona
UC Berkeley
zskatona@berkeley.edu

Marcus O. Painter
Saint Louis University
marc.painter@slu.edu

Panos N. Patatoukas 
UC Berkeley
panos@haas.berkeley.edu (corresponding author)

Jean Zeng
National University of Singapore
jeanzeng@nus.edu.sg

Abstract

We use the introduction of satellite coverage of major retailers to study the capital market implications of unequal access to big data. Satellite data enabled sophisticated investors with access to such data to formulate profitable trading strategies, especially by targeting the upcoming reports of retailers with bad news for the quarter. The introduction of satellite data led to more informed short-selling activity, less informed individual buying activity, and lower stock liquidity around the reports of retailers with satellite coverage. We conclude that unequal access to big data can increase information asymmetry among market participants without immediately enhancing price discovery.

I. Introduction

What are the implications of unequal access to big data for capital markets? Advancements in computational power, expanded data storage capacity, and faster

We thank RS Metrics and Orbital Insight for providing satellite imagery data of parking lot traffic, and Markit for providing securities lending market data. For helpful comments and discussions, we thank Hendrik Bessembinder (the editor), Chris Clifford, Zhi Da (the referee), Mike Farrell, Jill Fisch, Jillian Grennan, Kristine Hankins, Franz Hinzen (discussant), Russell Jame, Thomas Lee, Tamara Nefedova (discussant), Frank Partnoy, the PhD students at Berkeley Haas, former SEC Commissioner Robert Jackson, the DERA of the SEC, and the members of the University of Chicago Applied Math Club. We also thank seminar participants at the 2018 Miami Behavioral Finance Conference, the University of Florida, the University of Kentucky, the University of Missouri, the 2019 Future of Financial Information Conference at the Stockholm Business School, the 2019 Annual Meeting of the Financial Management Association, Harvard Business School, Kellogg School of Management, and Tulane University. We gratefully acknowledge financial support from the Fisher Center for Business Analytics at Berkeley Haas. This article is the product of merging two related papers, one of which had the current title and another that was titled “Un-Levelling the Playing Field: The Investment Value and Capital Market Consequences of Alternative Data.”

interconnection speeds have enabled access to large amounts of alternative, non-financial data that can inform investment decisions. However, access to big data is often only within the reach of sophisticated investors who can afford to incur the substantial costs of acquiring, processing, and integrating such data. These costs lead to unequal access to big data across Main Street and Wall Street investors.

One view of capital markets is that the introduction of big data could immediately enhance price discovery even if access is restricted to a few sophisticated investors. This view effectively assumes that stock prices instantaneously aggregate and disseminate value-relevant information embedded in data that would otherwise be inaccessible to all investors. This view goes back to Hayek's (1945) notion of the market as a mechanism for aggregating dispersed bits of knowledge in society and prices as a system for communicating all relevant information to every market participant. An alternative view of capital markets is that unequal access to big data leaves Main Street investors outside the "information loop" and creates trading opportunities for sophisticated investors without immediately enhancing stock price discovery.

Bringing all investors into the information loop has long been a key challenge for the SEC in its mission to serve and safeguard the interests of Main Street investors. Twenty-three years ago, this challenge was emphasized in the opening statement of former SEC Chair Arthur Levitt at the open meeting on Regulation Fair Disclosure (Reg FD) on Aug. 10, 2000: "...Like that neighborhood with gated entrances and tall fences, moving into the information loop is not always an option for many of America's small investors." Reg FD targeted the selective disclosure so that small investors have access to material nonpublic information, such as early warnings of earnings results, at the same time as Wall Street professionals. Fast forward to today, the rise of big data raises the question of whether another tall fence has been raised leaving individual investors outside the information loop.

We use the emergence of satellite imagery of parking lot traffic across major U.S. retailers to study how big data affects information asymmetry among market participants. The idea of measuring parking lot traffic to predict retailer performance had existed long before the introduction of satellite data. Sam Walton, the founder of Walmart, was known for routinely flying over parking lots so he could monitor store performance by counting cars (Walton and Huey (1992)). While anyone could count cars in a parking lot, advances in computer vision and the increased availability of satellite imagery has enabled daily tracking and processing of retailer parking lots at scale. This setting provides an important opportunity to study the capital market implications of unequal access to big data.

Due to the high acquisition and processing costs, access to satellite data is only within the reach of sophisticated investors, with select hedge funds being the typical clients of satellite data vendors. Satellite imagery is one of the most desirable data sources among hedge fund managers. However, financial analysts and mutual fund managers have not been widespread adopters citing various factors, including the cost and difficulty of accessing and quantifying the value of satellite data as well as the challenges of integrating such data with their internal resources (IHS Markit (2019)). These factors have been significant roadblocks in the path of democratizing access to satellite data for the general public.

Our primary data source is RS Metrics, the first vendor to introduce nearly real-time daily data feeds of store-level parking lot traffic signals extracted from

satellite imagery analysis in the U.S. market. The store-level data set includes 4.7 million daily observations across 67,078 unique store locations for 44 major U.S. retailers between 2011:Q1 and 2017:Q4. The daily feeds include information about parking lot capacity and utilization. From the daily store-level parking lot information, we develop enterprise-level measures of parking lot fill rates. Our sample of major retailers' accounts for a large fraction of the aggregate sales and market cap across U.S. listed firms in the same GICS industry.

We first validate the information content of satellite data for anticipating quarterly retailer performance. We focus on predicting quarterly earnings announcements because these are highly anticipated public events with significant price impacts. The evidence confirms that growth in parking lot fill rates is incrementally relevant for anticipating retailer performance for the quarter. Our results also show that the fundamental link between growth in parking lot fill rates and retailer performance is asymmetric. Specifically, we document that negative growth in parking lot utilization is an especially strong indicator of low fundamental performance for the quarter. This asymmetric fundamental link suggests that in-store foot traffic is more sensitive to decreases in parking lot utilization than to increases in parking lot utilization.

When information is costly to acquire, process, and integrate, standard theory predicts that it will not be immediately reflected in prices allowing informed investors to generate a competitive rate of return that is commensurate with the information acquisition costs (Grossman and Stiglitz (1980)). Put differently, the market cannot be perfectly efficient with respect to the information content of hard-to-access and hard-to-process big data but only "efficiently inefficient" conditional on the data costs (Pedersen (2015)). Under efficiently inefficient markets, investors with access to satellite data could profit by targeting the quarterly reports of retailers with satellite coverage. Indeed, a trading strategy that takes a long/short position in the stock of retailers that experience an abnormal increase/decrease in parking lot fill rates during the quarter generates abnormal returns around quarterly earnings announcements.

The portfolio returns from targeting retailers with satellite coverage are asymmetric on the long and the short side. We separate "good news" retailers that experience an abnormal increase in parking lot fill rates from "bad news" retailers that experience an abnormal decrease in parking lot fill rates during the quarter. The absolute magnitude of returns is nearly twice as large for the portfolio of bad news retailers relative to the good news portfolio. Over the 3-day window around quarterly earnings announcements, the good news portfolio outperforms the market by +1.6% while the bad news portfolio underperforms by nearly -3%, net of stock loan fees.

The distinct asymmetry in the long-short portfolio returns is consistent with our evidence that abnormal decreases in parking lot utilization are especially important for anticipating retailer performance. The return asymmetry implies that satellite data is particularly relevant for short sellers interested in targeting retailers with bad news for the quarter. Indeed, using daily stock lending data from Markit, we find evidence of informed short-selling activity leading to the quarterly reports of retailers with satellite coverage. Focusing on bad news retailers with abnormal decreases in parking lot fill rates during the quarter, our evidence shows a substantial increase in the lender quantity on loan over the 5 trading days leading to their

quarterly reports. On the other side, we do not observe pre-earnings announcement changes in short-selling activity for good news retailers.

Notwithstanding evidence of informed short selling leading to the quarterly reports of bad news retailers, individual investors cannot “piggyback” on the buildup of short-selling activity. The reason is that the general public has access to short-interest data only twice per month and only with a significant delay. In fact, our analysis of pre-announcement effects shows only limited evidence of stock price discovery prior to the quarterly report, with most of the trading gains being realized on the quarterly earnings announcement days.

While short sellers are actively targeting the quarterly reports of retailers with bad news for the quarter, we find evidence that individual investors are net buyers of the stock of such retailers. The dynamics of short-selling activity and individual trading closely mirror each other around the release of bad news for the quarter. To estimate the impact of the introduction of satellite data on trading activity, we implement a difference-in-differences (DID) design. The DID design compares the treated group of retailers with satellite coverage to the matched control group of retailers with no satellite coverage in the periods before and after the initiation of satellite coverage.

The DID results show that the ability of short sellers to profit from targeting retailers with bad news for the quarter has improved after the introduction of satellite data for the group of retailers with satellite coverage. In contrast, our evidence also shows that after the introduction of satellite data, individual investors’ trades, especially individual investor buys, have become less informative with respect to anticipating retailer news for the quarter. The opposing impact of the introduction of satellite data on short-selling activity versus individual trading activity implies that differential access to alternative data can lead to an increase in information asymmetry among market participants. Consistent with this implication, the DID results also show that the introduction of satellite data led to a decrease in stock liquidity around the quarterly reports of retailers with satellite coverage, as indicated by an increase in their bid–ask spread and price impact.

In our final set of tests, we evaluate the impact of the introduction of satellite data on the speed of price discovery. Timelier price discovery would imply that stock prices impound more information prior to a quarterly report, so that the reaction to the public announcement itself is muted. The DID results show that on average the introduction of satellite data had no detectable effect on the speed of price discovery. A closer investigation reveals that this null result masks a gradual, long-term impact on the price discovery of retailers with satellite coverage that gets attenuated when averaging across years in the post-treatment period. The evidence suggests that the introduction of satellite data did not immediately enhance price discovery; instead, its impact on price informativeness emerged over time with the increased adoption of alternative data.

Our article contributes to research on the role of big data in capital markets. Froot, Kang, Ozik, and Sadka (2017) use proprietary data of consumer activity and find that managers distort their disclosures in the presence of insider trading opportunities. Zhu (2019) finds that the introduction of big data decreased information asymmetry between corporate insiders and sophisticated investors. Dessaint, Foucault, and Frésard (2021) propose that alternative data are informative

about a firm's short-term prospects but not as informative about a firm's long-term prospects, though Chang and Da (2022) provide evidence that big data has long-lasting predictive power. Building on our study, Gerken and Painter (2019) and Kang, Stice-Lawrence, and Wong (2021) study how institutional investors and sell-side analysts react to changes in the local performance of retail firms, and Cao, Jiang, Wang, and Yang (2021) find that satellite data enabled sell-side analysts to forecast earnings more accurately than an AI algorithm. At the macro level, Mukherjee, Panayotov, and Shon (2021) provide evidence that crude oil price responds more to government announcements of oil inventories in cloudy weeks, when satellite estimates of U.S. oil inventories are less accurate, compared to clear weeks, when such estimates are likely to be more accurate.

Different from prior work, we zero in on the effect of big data on informed short-selling activity, uninformed individual trading, stock liquidity, and price discovery leading to quarterly earnings announcements. Our article provides new evidence that unequal access to big data can actually increase information asymmetry among market participants without immediately enhancing stock price discovery. More broadly, our article contributes to research on the impact of technology and data abundance on capital markets proposing that advances in technology and alternative data have nuanced effects on information asymmetry, price discovery, and investor participation (e.g., Banerjee, Davis, and Gondhi (2018), Dugast and Foucault ((2018), (2021)), Weller (2018), and Grennan and Michaely (2021)).

A relevant policy question emanating from our evidence on the impact of unequal access to big data is what the benefits and costs of regulation aimed at the public disclosure of more precise and timely information about short positions would be. This question is a timely since on Oct. 13, 2023, the SEC adopted new Rule 13f-2 and new Form SHO aimed at providing greater transparency through the publication of short sale-related data to investors and other market participants (Rel. No. 34-98738). While regulatory policies aimed at the public disclosure of short positions can allow the general public to learn from informed traders, such policies may also cause distortions in capital markets (e.g., Jones, Reed, and Waller (2016), Jank, Roling, and Smajlbegovic (2021)), and Ahn, Bushman, and Patatoukas (2023).

II. Measuring Parking Lot Traffic from Outer Space

A. Satellite Imagery Data of Parking Lot Traffic

Our primary source of parking lot data is RS Metrics, the first U.S. data vendor to introduce nearly real-time parking lot traffic signals from satellite imagery, starting in 2011:Q1.¹ Appendix A in the Supplementary Material provides the

¹While RS Metrics is the first data vendor to sell domestic parking lot signals beginning in 2011:Q1, there are other competing data vendors sourcing imagery from the same satellites with Orbital Insight being the most prominent competitor in the U.S. Orbital Insight, however, started selling parking lot traffic signals to investors beginning in mid-2015 (i.e., more than 4 years after investors could subscribe to RS Metrics' data feeds). In an additional analysis, we merge store-level parking lot data from RS Metrics and Orbital Insight and find that investors with access to both vendors could formulate even

background on remote sensing technology. The data consists of daily store-level information about parking lot capacity and utilization across major U.S. retailers. With respect to the cost of accessing satellite imagery data, data vendors privately negotiate the price depending on the timeliness and extent of access. Whereas the estimated cost of accessing the data is in the range of hundreds of thousands of dollars per year, this estimate does not account for the significant costs of processing and integrating the data.

The measurement of parking lot traffic is subject to error. First, satellite coverage is available only for a subset of a retailer's stores. One reason for partial coverage is that the cameras have to be pointed in a given direction for a certain store and there is only limited capacity allotted to each satellite user. Another reason is that not all parking lots are visible from outer space (e.g., underground, and multi-story lots). In addition, satellite coverage is restricted to domestic store locations. Second, the satellite's orbit is designed in a way that it passes through a given latitude at the same local time between late morning and early afternoon for most of the continental U.S., which captures only a snapshot of total parking lot traffic during the day. Third, though the resolution of satellite imagery has improved over time, clouds, haze, trees, shadows, and other environmental factors obscure the imagery. RS Metrics processes satellite imagery using a combination of a software for automated counts and human analysts for verifying the counts.

B. Measuring Parking Lot Traffic at the Individual Store Level

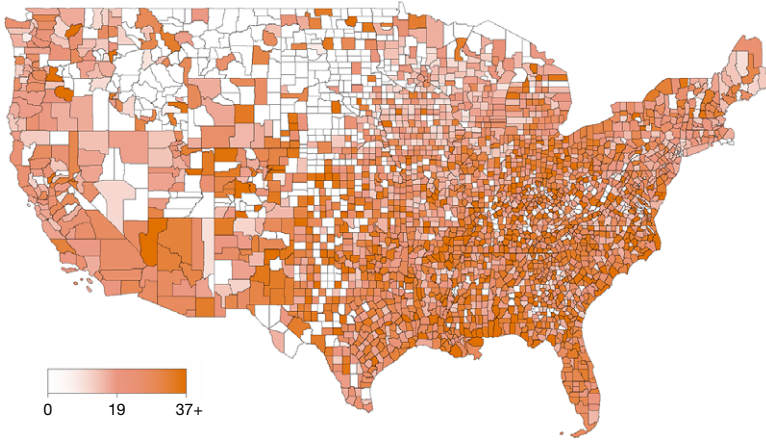
Our sample starts in 2011:Q1 because this is the first quarter for which RS Metrics started selling satellite imagery data. Our sample ends in 2017:Q4 because this is the last quarter for which we obtained satellite data from RS Metrics per our data service agreement. Our sample includes 4.7 million daily observations across a total of 67,078 unique store locations for the 44 major U.S. retailers with satellite coverage. The key information available from the processed satellite imagery is the daily number of cars parked in an individual store parking lot; denoted by $CARS_{ijd}$, along with the total number of available parking spaces; denoted by $SPACES_{ijd}$, where i indicates the retailer, j indicates the individual store location, and d indicates the day of the satellite imagery.

Table A1 in the Supplementary Material reports information about the store count and satellite store coverage for each of the 44 companies in our sample along with the starting date of satellite coverage. The cross-sectional average store count is 2,412 with satellite coverage available for 58% of the individual store locations. We organize our sample using 6-digit Global Industry Classification Standard (GICS) codes. The most represented group in our sample is specialty stores with 21 retailers, including Walmart Inc. and Target Corporation. Figure 1 maps the geographical coverage of satellite imagery and shows that our store-level data covers 2,571 counties representing 98% of the U.S. population.

more profitable strategies targeting earnings announcements. The complementarities across data vendors can help investors extract more accurate signals.

FIGURE 1
Geographical Coverage of Satellite Imagery

Figure 1 presents the number of individual store locations with satellite coverage per 100,000 residents across counties in the U.S. The underlying data covers 67,078 individual store locations across 2,571 counties covering 98% of the U.S. population. For each county, we compute the number of individual store locations with satellite coverage per 100,000 residents. Across counties, the mean (median) store count per 100,000 residents is 18.11 (18.61) stores, with standard deviation of 12.48 and interquartile range from 9.55 to 26.55. The heat map shows that satellite coverage is extensive not only in densely populated areas, but also in more rural counties with the exception of the most sparsely populated ones. In fact, the mean (median) population of counties with no coverage in our data is 7,537 (5,705), while that of counties with coverage is 117,725 (35,767). The color spectrum across counties is proportionately dark to the number of store coverage per capita ranging from white to dark orange. Across counties, the median store count per 100,000 residents is 18.6 stores with interquartile range from 9.6 to 26.6.



C. Measuring Parking Lot Traffic at the Enterprise Level

We start with the daily data for each individual store location j during quarter q and compute the average number of cars parked during the quarter ($CARS_{ijq}$) as well as the average number of parking lot spaces available at each store location during the quarter ($SPACES_{ijq}$). Due to seasonal effects in quarterly data, we focus on year-over-year (YoY) comparisons rather than sequential comparisons (i.e., we compare quarter q to quarter $q - 4$). To ensure YoY comparability, we restrict our attention to individual store locations with satellite imagery in both quarter q and quarter $q - 4$. The same-store comparisons control for YoY growth in parking lot capacity due to acquisitions and the opening of new stores. Figure 2 provides an illustrative example of satellite imagery for Target Corporation, the department store company.

Our final sample includes 3.4 million daily observations across 53,647 unique store locations for 44 major U.S. retailers covered from 2011:Q1 to 2017:Q4. From the daily store-level data, we compile a panel of 650 firm-quarter observations of enterprise-level parking lot fill rates. For each retailer-quarter, we sum up across individual store locations with YoY satellite coverage to obtain the aggregate parking lot traffic; $CARS_{iq}$, and the aggregate parking lot space; $SPACES_{iq}$. We calculate the enterprise-level parking lot fill rate – our measure of overall parking lot utilization – as the ratio of aggregate parking lot traffic divided by aggregate parking

$$\text{lot space: } FLRT_{iq} = \frac{\sum_{j=1}^J CARS_{ijq}}{\sum_{j=1}^J SPACES_{ijq}} = \frac{CARS_{iq}}{SPACES_{iq}}.$$

FIGURE 2
Illustrative Example of Satellite Imagery

Figure 2 illustrates the measurement of key variables using satellite imagery data for Target Corporation, the department store company. The satellite image is for the Target store located at 4500 Macdonald Ave, Richmond CA 94805. The image was captured on Sept. 19, 2016, at 11:03AM. The processed image indicates the number of cars present within a fixed area of parking lot spaces that RS Metrics assigns to each store. The yellow line outlines the boundary of the parking lot associated with Target and the red dots indicate the occupied parking lot spaces. The parking lot spaces assigned to each store do not change over time unless the company renovates the parking lot. At the time of the satellite image, RS Metrics reports 540 parking lot spaces with 146 of them filled. The parking lot spaces on the bottom right of this Target store are excluded because they may represent employee parking. As a general rule for any individual store location, RS Metrics defines the "most likely parking area" for customers and keeps that parking lot boundary relatively fixed over time so that the variability in the data comes from the number of cars parked at any time. Starting with the granular parking lot data for Target Corporation in 2016:Q3, we identify 1,210 individual store locations across the U.S. with year-over-year satellite coverage (i.e., coverage in both 2016:Q3 and 2015:Q3). We calculate the average parking lot size and parking lot traffic per Target store during the quarter, and we sum across stores to obtain the enterprise-level information. For 2016:Q3 across the 1,210 Target store locations with year-over-year satellite coverage, the aggregate parking lot traffic is 156,977 while the aggregate parking lot space is 595,340. It follows that the parking lot fill rate for Target Corporation in 2016:Q3 is 26.37%. Repeating the steps for 2015:Q3, we find a fill rate of 26.94%. Hence, the year-over-year growth rate in the fill rate is -2.14% .



The key variable of interest is the YoY growth in same-store parking lot fill rates: $\Delta FLRT_{iq} = \frac{FLRT_{iq} - FLRT_{iq-4}}{FLRT_{iq-4}}$. We note that the variability in same-store parking lot fill rates is primarily due to variability in parking lot traffic rather than parking lot capacity. This is because parking lot capacity at the individual store-level is very steady from 1 year to the next. Indeed, growth in same-store parking lot fill rates is 99% correlated with growth in same-store car traffic.

D. Descriptive Statistics

Panel A of Table 1 reports the empirical distributions of key variables. Appendix B in the Supplementary Material provides key variable definitions. The sample includes 650 firm-quarter observations across 44 major U.S. retailers from 2011:Q1 to 2017:Q4. The parking lot fill rate has a mean (median) value of 29.8% (26.8%) with a standard deviation of 9.9% and an interquartile range of 22.9% to 35.3%. The distribution of the YoY growth in parking lot utilization ($\Delta FLRT_{iq}$) is centered at -0.7% and exhibits substantial variation with a standard deviation of 4.9% and an interquartile range of -3.4% to 1.8%. The pairwise correlations in Panel B of Table 1 provide preliminary evidence that $\Delta FLRT_{iq}$ is relevant for anticipating the market reaction to quarterly earnings announcements.

TABLE 1
Descriptive Analysis

Table 1 presents descriptive statistics. Panel A reports the empirical distributions of key variables. Panel B reports Pearson (Spearman) pairwise correlations below (above) the main diagonal. Panel C reports the distribution of the 44 U.S. companies in our sample across their 6-digit GICS industries and their contribution to aggregate sales and market value separately for each industry. *** indicates statistical significance at the 1% level using 2-tailed tests. The sample includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4.

Panel A. Empirical Distributions

	Mean	Std.Dev.	Min	p25	p50	p75	Max
ΔSSS_{iq}	0.013	0.057	-0.317	-0.012	0.016	0.043	0.240
$\Delta FLRT_{iq}$	0.298	0.099	0.131	0.229	0.268	0.353	0.604
$\Delta FLRT_{iq}$	-0.007	0.049	-0.295	-0.034	-0.007	0.018	0.415
$QRET_{iq}$	-0.026	0.163	-0.624	-0.120	-0.016	0.066	0.773
$EARET_{iq}$	-0.004	0.097	-0.403	-0.054	-0.002	0.048	0.472

Panel B. Pairwise Correlations

	1	2	3	4
(1) ΔSSS_{iq}		0.371***	0.336***	0.203***
(2) $\Delta FLRT_{iq}$	0.383***		0.047	0.130***
(3) $QRET_{iq}$	0.280***	0.032		0.021
(4) $EARET_{iq}$	0.174***	0.123***	-0.017	

Panel C. Industry Breakdown

Industry Name	GICS code	# Of firms	% Sales	% Mkt Cap
Multiline retail	255,030	10	85	77
Specialty retail	255,040	21	53	70
Food and staples retailing	301,010	6	57	52
Hotels, restaurants and leisure	253,010	6	9	12
Chemicals	151,010	1	4	4

Panel C of Table 1 reports the sample distribution across 6-digit GICS industries along with their contribution to the aggregate sales and market value separately for each industry. The evidence shows that our sample of major retailers accounts for a large fraction of the aggregate sales and the aggregate market value across U.S. listed companies operating in the same industry. To illustrate, our sample includes 10 multiline retail companies, which account for as much as 85% of the sales and 77% of the market value of the U.S. listed firms in the same GICS code.

III. Capital Market Consequences

A. Forward-Looking Content of Satellite Imagery Data

We first validate the relevance of satellite imagery of parking lot fill rates for anticipating retailer performance. The idea is that variation in parking lot utilization should be correlated with shopper conversion across stores. Higher YoY growth in same-store parking lot utilization should indicate higher closing rates and, therefore, higher same-store sales growth. We obtain quarterly data on same-store sales (SSS_{iq}) from FactSet Fundamentals. We focus on the domestic portion of sales because satellite imagery covers only individual stores located in the U.S. We measure YoY growth in the domestic portion of same-store sales as

$\Delta SSS_{iq} = (SSS_{iq} - SSS_{iq-4}) / SSS_{iq-4}$. We focus on YoY growth rather than sequential growth due to seasonal effects in retailer performance.² Our first set of tests are based on regression models of the following form:

$$(1) \quad \Delta SSS_{iq} = \alpha + \beta_1 \Delta FLRT_{iq} + \beta_2 \Delta SSS_{iq-1} + \beta_3 QRET_{iq} + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}.$$

The dependent variable is domestic same-store sales growth (ΔSSS_{iq}) and the set of right-hand-side predictors includes same-store growth in parking lot utilization ($\Delta FLRT_{iq}$), lagged same-store sales growth (ΔSSS_{iq-1}), and the stock return from the beginning to the end of quarter q ($QRET_{iq}$).³ The vector C_{iq} spans an array of time-varying fundamental attributes, including log market capitalization, Tobin's Q, institutional ownership, and indicators for Big-4 auditors, acquisitions, restructurings, asset write-downs, and impairments. The model includes firm fixed effects (θ_i) to control for firm-specific time-invariant factors and quarter fixed effects (δ_q) to control for aggregate time-varying factors. To ease the interpretation of the estimates, we report regression results using the standardized z -values of the continuous predictors. The standardized regression coefficients measure changes in standard deviation units, which allows us to easily compare the relative importance of each predictor.

Panel A of Table 2 reports regression results for equation (1). The evidence confirms that $\Delta FLRT_{iq}$ is an incrementally relevant predictor of retailer performance after accounting for autocorrelation in retailer growth as well as forward-looking information embedded in the quarterly stock return. Put differently, the predictive content of YoY growth in parking lot utilization contains information that is not subsumed by signals as easily accessible as the past realization of same-store sales growth and the quarterly stock return. A 1-standard-deviation increase in $\Delta FLRT_{iq}$ is associated with a 0.8% increase in ΔSSS_{iq} representing a 61.5% deviation from the mean. Comparing model specifications, we observe that the magnitude of the estimated coefficient on $\Delta FLRT_{iq}$ is not sensitive to the inclusion of time-varying firm characteristics and fixed effects.

Panel B of Table 2 separates negative from positive values of $\Delta FLRT_{iq}$ and provides evidence of asymmetry in the fundamental implications of negative versus positive growth in parking lot utilization. Across specifications, the magnitude of the estimated coefficient on $\Delta FLRT_{iq}^-$ is nearly 3 times that on $\Delta FLRT_{iq}^+$. Put differently, the evidence underscores that negative growth in parking lot utilization is an especially strong indicator of low fundamental performance for the current quarter. The asymmetric effect of $\Delta FLRT_{iq}$ on ΔSSS_{iq} suggests that in-store foot traffic is more sensitive to decreases in parking lot utilization than to increases in parking lot utilization.

²For retailers that are growing by opening new stores, same-store comparisons allow investors to differentiate between growth that comes from new stores, and growth from improved operations at existing store locations. Table A2 in the Supplementary Material shows that our inferences are unchanged when we replace growth in same-store parking lot fill rates with overall growth without conditioning on same-store comparisons. The correlation between the two growth measures is 89%.

³Table A3 in the Supplementary Material shows that our inferences are unchanged when we replace the stock return cumulated from the beginning to the end of quarter q ($QRET_{iq}$), with the stock return cumulated from the beginning of quarter q to 2 days before the quarterly earnings announcement ($QRET_{iq}^*$).

TABLE 2
Forward-Looking Content of Satellite Imagery Data

Table 2 provides evidence that growth in same-store parking lot fill rates predicts growth in same-store retailer sales. Panel A reports results from the baseline linear regression model in equation (1). Panel B reports results from the alternative specification that allows for a different coefficient on negative and positive values of growth in same-store parking lot fill rates. We report regression results using the standardized z-values of the continuous predictors. The sample includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. We report t-statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, using 2-tailed tests.

Panel A. Baseline Specification

	Dependent Variable: $\Delta SSS_{i,q}$		
	1	2	3
$\Delta FLRT_{i,q}$	0.008** (3.46)	0.008*** (3.75)	0.008*** (4.41)
$\Delta SSS_{i,q-1}$	0.044*** (15.39)	0.043*** (16.53)	0.039*** (12.59)
$QRET_{i,q}$	-	0.007*** (3.87)	0.007*** (7.64)
Characteristic controls	No	No	Yes
Firm fixed effects	No	No	Yes
Quarter fixed effects	No	No	Yes
Adj. R^2	69.1%	70.6%	71.9%
No. of obs.	650	650	650

Panel B. Alternative Specification

	Dependent Variable: $\Delta SSS_{i,q}$		
	1	2	3
$\Delta FLRT_{i,q}^-$	0.017*** (5.57)	0.016*** (4.95)	0.015** (3.39)
$\Delta FLRT_{i,q}^+$	0.006* (2.33)	0.006** (2.79)	0.005** (2.55)
$\Delta SSS_{i,q-1}$	0.043*** (18.63)	0.042*** (19.52)	0.038*** (14.02)
$QRET_{i,q}$	-	0.007*** (3.79)	0.006*** (7.92)
Characteristic controls	No	No	Yes
Firm fixed effects	No	No	Yes
Quarter fixed effects	No	No	Yes
Adj. R^2	70.0%	71.3%	72.3%
No. of obs.	650	650	650

Together, the evidence suggests that satellite data is incrementally relevant for anticipating retailer performance, especially for retailers experiencing a decrease in parking lot utilization. If a subset of sophisticated investors is able to profit from trading based on satellite data, the introduction of satellite data could lead to an increase in information asymmetry between investors who can access and process the data and those who cannot. Next, we examine whether investors with access to satellite data could formulate a trading strategy targeting the quarterly reports of retailers with satellite coverage.

B. Formulating a Trading Strategy Using Satellite Imagery Data

Panel A of Table 3 reports the results from a trading strategy that buys/short sells retailers with same-store parking lot fill rate growth in the top/bottom quartile of the cross-sectional distribution of $\Delta FLRT_{i,q}$. We report raw returns, market-adjusted returns, as well as size and book-to-market factor-adjusted returns, or abnormal returns, cumulated over the 3-day window centered on the quarterly

TABLE 3
Formulating a Trading Strategy Using Satellite Imagery Data

Table 3 reports returns from a trading strategy that buys (short sells) retailers with same-store parking lot fill rate growth in the top (bottom) quartile of the cross-sectional distribution of ΔFLRT_{it} . We report raw, market-adjusted, and size and book-to-market factor-adjusted returns over the 3-day window centered on quarterly earnings announcements. We use the value-weighted CRSP index including distributions when calculating market-adjusted returns. We use the portfolio data from Kenneth French's website to calculate factor-adjusted returns. To generate the cross-sectional quartile cutoff values of ΔFLRT_{it} , we consider retailers with fiscal quarters ending within the last 3 months. Panel A (Panel B) reports results before (after) adjusting the short-sell portfolio returns for accumulated stock loan fees. The sample includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, using 2-tailed tests.

Panel A. Portfolio Performance Before Stock Loan Fees

	Portfolio Returns Before Stock Loan Fees		
	Raw Returns	Market Adjusted	Factor Adjusted
Sell portfolio	-2.82%*** (-2.90)	-3.01%*** (-3.13)	-3.10%*** (-3.25)
Buy portfolio	1.78%** (2.38)	1.63%** (2.17)	1.66%** (2.22)
Buy-minus-sell	4.60%*** (3.75)	4.64%*** (3.80)	4.76%*** (3.93)

Panel B. Portfolio Performance After Stock Loan Fees

	Portfolio Returns After Stock Loan Fees		
	Raw Returns	Market Adjusted	Factor Adjusted
Sell portfolio	-2.79%*** (-2.87)	-2.98%*** (-3.10)	-3.07%*** (-3.22)
Buy portfolio	1.78%** (2.38)	1.63%** (2.17)	1.66%** (2.22)
Buy-minus-sell	4.57%*** (3.73)	4.62%*** (3.78)	4.73%*** (3.90)

earnings announcement. To generate the cutoff values of ΔFLRT_{it} , we consider retailers with fiscal quarters ending within the last 3 months. This approach allows us to generate cross-sectional cutoff values on a rolling basis, thereby, allowing for time-series variability in the empirical distribution of ΔFLRT_{it} .

The evidence shows that around quarterly earnings announcements, the short-sell portfolio of retailers with abnormal decreases in parking lot fill rates underperforms the market by -3.10% while the buy portfolio of retailers with abnormal increases in parking lot fill rates outperforms the market by 1.66%. In terms of abnormal returns, the spread between the buy and sell portfolios is 4.76%, which is statistically significant and economically important. Though we find significant abnormal returns for both portfolios, the evidence shows that the absolute magnitude of returns is nearly twice as large for the short-sell portfolio relative to the buy portfolio. The marked asymmetry in returns is consistent with our findings of an asymmetric fundamental link in Panel B of Table 2.

Importantly, the documented asymmetries imply that satellite data is especially relevant for short sellers interested in targeting the earnings announcements of retailers with bad news for the quarter. To evaluate the performance net of shorting fees, we obtain daily data on stock lending market conditions from Markit. Markit aggregates information from institutional lenders that collectively account for most of the lendable inventory of shares in the U.S. Panel B of Table 3 reports results after adjusting returns for the cumulative stock loan fees. We find that the

stock loan fees are less than 1 basis point per day. Therefore, the portfolio returns remain significant after accounting for the direct cost of short selling.

Graph A of Figure 3 reports the cumulative abnormal returns separately for the top and bottom ΔFLRT_{iq} portfolios over the ± 10 trading-day window centered on the earnings announcement (i.e., day 0). The green dashed line presents the performance of the buy portfolio of retailers with abnormal increases in parking lot fill rates. The red dotted (solid) line presents the performance of the short-sell portfolio after (before) stock loan fees. The evidence shows that there is only limited pre-announcement activity and the portfolio returns are effectively realized at the time of the earnings announcement. A key implication is that investors with access to satellite data could get ahead of the market and formulate a trading strategy anticipating the market reaction to quarterly earnings announcements. Again, the

FIGURE 3
Formulating a Trading Strategy Using Satellite Imagery Data

Figure 3 presents the cumulative factor-adjusted return from a strategy that buys (short sells) retailers with same-store parking lot fill rate growth in the top (bottom) quartile of the cross-sectional distribution of ΔFLRT_{iq} . The measurement window is from 10 trading days before to 10 trading days after the quarterly earnings announcement (day 0). In Graph A, the green dashed (red solid) line presents the performance of the portfolio that buys (short sells) retailers in the top (bottom) quartile portfolio of ΔFLRT_{iq} , while the red dotted line presents the performance of the short-sell portfolio net of stock loan fees. In Graph B, we present the daily abnormal share turnover for the buy and short-sell portfolios. We measure share turnover as the daily trading volume divided by the number of shares outstanding. To obtain the abnormal values of share turnover, we adjust the daily values of share turnover during the event window for the daily share turnover prior to the event window. The sample includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4.



evidence highlights the asymmetry in the performance of the bad news portfolio relative to the good news portfolio.

Graph B of Figure 3 presents daily values of abnormal share turnover for the buy and short-sell portfolios over the ± 10 trading-day window centered on day 0. We measure share turnover as the daily trading volume divided by the number of shares outstanding. To obtain the abnormal values of share turnover, we adjust the daily values of share turnover during the event window for the daily share turnover prior to the event window. We observe that abnormal share turnover is close to 0 prior to the earnings announcement, which is consistent with the lack of abnormal pre-announcement stock returns. Additionally, the significant spike in abnormal share turnover around day 0 suggests that most of the price discovery happens on the announcement day.

C. Informed Short-Selling Activity and Uninformed Individual Trading

The evidence thus far shows that investors with access to satellite data could get ahead of the market and formulate trading strategies targeting quarterly earnings announcements. The strategy works on both the long and the short side, though the returns are especially pronounced from targeting bad news retailers with abnormal decreases in parking lot fill rates. This distinct asymmetry implies that satellite data is especially relevant for short sellers. In what follows, we use daily data on lender quantity on loan from Markit to provide direct evidence of informed short-selling activity in the stock lending market.⁴

Graph A of Figure 4 presents the cumulative change of lender quantity on loan as a percentage of shares outstanding separately for the top and bottom ΔFLRT_{iq} portfolios. The evidence is consistent with informed short-selling activity prior to the earnings announcement. Focusing on the bottom ΔFLRT_{iq} portfolio (red solid line), we find evidence of a sharp increase in the lender quantity on loan starting 5 trading days prior to the earnings announcement. On the other side, we do not find evidence of significant changes in short-selling activity for the top ΔFLRT_{iq} portfolio (green dashed line).

We note that selling activity accounts for a relatively small fraction of overall trading around the quarterly earnings announcements. This is not surprising since short selling is comparatively risky, with no theoretical limit on how much the short seller might lose on their position if the stock prices rapidly rise. Importantly, the comparison of short-selling activity to overall trading volume highlights that the signals embedded in the activity of informed short sellers can get drowned out by noise. Theory predicts that noise in trading can camouflage the signals of informed traders, which can lead to informed traders reaping abnormal profits without an immediate impact on price discovery (e.g., Verrecchia (1982)).

⁴Markit records stock lending activity when it becomes known to the market (i.e., as of the settlement date). Up until 2017, the settlement date was the trade date plus 3 trading days. After Sept. 5, 2017, the SEC shortened the standard settlement cycle from 3 trading days after the trade date to 2 trading days (Rel. No. 34-80295). To match stock lending activity to the occurrence of an underlying short sale, we follow prior research and account for the trade settlement period by shifting stock loan transactions back by 2 or 3 trading days (e.g., Ahn and Patatoukas (2022)).

FIGURE 4

Informed Short-Selling Activity and Uninformed Individual Investor Trading

Graph A of Figure 4 presents the cumulative changes in the lender quantity on loan for the portfolio that buys (short sells) retailers in the top (bottom) quartile portfolio of $\Delta FLRT_{iq}$. To generate cross-sectional quartile cutoff values of $\Delta FLRT_{iq}$, we consider retailers with fiscal quarters ending within the last 3 months. The measurement window is from 10 trading days before to 10 trading days after the quarterly earnings announcement day (day 0). The green dashed (red solid) line presents the cumulative change in lender quantity on loan as a percentage of the number of shares outstanding for the portfolio that buys (short sells) retailers in the top (bottom) quartile portfolio of $\Delta FLRT_{iq}$. Graph B presents the cumulative daily order imbalance of individual investors in the top (bottom) quartile portfolio of $\Delta FLRT_{iq}$. We measure individual order imbalance as the total individual investor-initiated buys minus the total individual initiated sells divided by the total number of shares outstanding. The sample includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4.



Relatedly, individual investors cannot “piggyback” on the buildup of informed short-selling activity leading to the quarterly reports of bad news retailers. The reason is that under the current disclosure regime (FINRA Rule 4560), the general public has access to short-interest data only twice per month and only with a significant delay. The lack of timely disclosures of short positions together with the fact the short selling accounts for a relatively small fraction of overall trading activity around earnings announcements help reconcile the seeming disconnect between the buildup of short-selling activity and the lack of abnormal pre-announcement stock returns.

Next, we use the algorithm of Boehmer, Jones, Zhang, and Zhang (2021), henceforth BJZZ algorithm, to measure individual investor buy and sell trades. To measure individual order flow, we first identify off-exchange trades in the Trade and Quote (TAQ) data (exchange code “D”). Then, we identify trades as individual buys (sells) if the trade took place at a price just below (above) a round penny.

We measure individual order imbalance as the individual buys minus the individual sells divided by the total number of shares outstanding.

Graph B of Figure 4 presents the cumulative change in individual order imbalance for the top and bottom ΔFLRT_{iq} portfolios. The evidence is consistent with uninformed individual trading around earnings announcements. Specifically, the individual order imbalance for retailers in the short-sell portfolio increases significantly prior to the earnings announcement while the individual order imbalance for the buy portfolio remains near 0. Together the evidence shows that the dynamics of short-selling activity and individual trading mirror each other. As short sellers actively target retailers with bad news for the quarter, individual investors are net buyers of such retailers. Next, we present evidence that the ability of short sellers to profit from targeting quarterly reports has increased after the introduction of satellite data for the group of retailers with satellite coverage.

D. Identifying the Effect on Short-Selling Activity

What is the effect of the introduction of satellite data on short-selling activity? An increase in the informativeness of short-selling activity would imply that short-sellers' ability to anticipate bad news increases after the introduction of satellite coverage. To estimate the effect of satellite coverage on the informativeness of short-selling activity, we implement a difference-in-differences (DID) design using the following model:

$$(2) \quad \text{EARET}_{iq} = \alpha + \beta_1 \text{POST}_{iq} + \beta_2 \text{TREAT}_{iq} + \beta_3 \text{POST}_{iq} \times \text{TREAT}_{iq} + \beta_4 \Delta\text{SHORT}_{iq} + \beta_5 \text{POST}_{iq} \times \Delta\text{SHORT}_{iq} + \beta_6 \text{TREAT}_{iq} \times \Delta\text{SHORT}_{iq} + \beta_7 \text{POST}_{iq} \times \text{TREAT}_{iq} \times \Delta\text{SHORT}_{iq} + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}.$$

The dependent variable EARET_{iq} in equation (2) is the cumulative abnormal return from 1 trading day before to one trading week after the quarterly earnings announcement. Turning to the right-hand-side variables, POST_{iq} is an indicator variable that takes the value 1 after the initiation of satellite coverage, TREAT_{iq} is an indicator variable that takes the value 1 for retailers in the treated group, and ΔSHORT_{iq} is either the cumulative change in the lender quantity on loan, $\Delta\text{SHORT}_{iq}^{\text{Dem}}$, or the cumulative change in the available supply of lendable shares, $\Delta\text{SHORT}_{iq}^{\text{Sup}}$, both expressed as a percentage of shares outstanding and measured from the end of the quarter to 2 days before the earnings announcement.

The DID design compares the group of covered retailers to the matched control group. For each covered firm, we use a symmetric event window before and after the initiation of satellite coverage. To construct the matched control group, we use the FactSet Revere database to identify all named competitors, including those reported by the target company itself or by the competitors.⁵ Then, we zero in on

⁵FactSet Revere provides the most comprehensive coverage of firm relationships that is currently available (e.g., Gofman, Segal, and Wu (2020)). FactSet analysts monitor the relationships and collect information from firms' annual reports, press releases, investor presentations, and investor relation websites. FactSet Revere's named competitors are widely used for comparable company analysis.

named competitors that operate in the same 6-digit GICS industry and do not have satellite coverage. Our choice of a fine industry definition ensures higher comparability across matched pairs. Lastly, we restrict the control group to include the closest-named competitors in terms of market capitalization by minimizing the absolute distance across all matched pairs.⁶

Our procedure identifies an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit GICS industry and do not have satellite coverage. The sample of treated and matched-control firms includes 4,420 observations with pre- and post-treatment coverage. Across DID analyses, the number of observations varies depending on the availability of each test variable. Table A4 in the Supplementary Material reports that retailers with satellite coverage tend to have a higher valuation, more institutional ownership, and are more likely to be audited by a Big-4 audit firm. To control for cross-sectional differences, the vector of time-variant firm characteristics (C_{iq}) includes log market cap, Tobin's Q, institutional ownership, the Big-4 auditor indicator, as well as indicators for material corporate events, including acquisitions, restructurings, write-downs, and impairments.

The coefficient on the triple interaction $POST_{iq} \times TREAT_{iq} \times \Delta SHORT_{iq}^{Dem}$ captures the change in the differential predictive ability of short-selling demand for earnings announcement returns across treated and matched control groups after the introduction of satellite data. An increase in the informativeness of short-selling demand would imply that $\Delta SHORT_{iq}^{Dem}$ is more negatively related to earnings announcement returns in the post period (i.e., $\beta_7 < 0$).

Table 4 reports DID regression results for equation (2). The significantly negative coefficient on the triple interaction $POST_{iq} \times TREAT_{iq} \times \Delta SHORT_{iq}^{Dem}$ in column 1 is consistent with an increase in the informativeness of short-selling demand for retailers with satellite coverage after the introduction of satellite data. Focusing on the change in the differential predictive ability of short-selling demand for earnings announcement returns, a 1-standard-deviation increase in pre-announcement lender quantity on loan is associated with a 2.4% lower earnings announcement return for treated retailers relative to the matched control group after the introduction of satellite data. In absolute terms, this magnitude is 5.8 times the value of the average earnings announcement return. Turning to column 2, we observe that the coefficient on the triple interaction $POST_{iq} \times TREAT_{iq} \times \Delta SHORT_{iq}^{Sup}$ is indistinguishable from 0, which implies that there is no detectable change in the informativeness of the supply of lendable shares. This result suggests that evidence of an increase in the informativeness of short demand after the introduction of satellite data is not confounded by overlapping changes in the informativeness of short supply.

A key assumption of the DID estimation is that in the absence of treatment, the average change in outcomes would have been the same for both the treatment and

In additional analysis, we find similar results when we match target companies to the closest industry peers without conditioning the matched-control group to include named competitors.

⁶We note that our matching does not use early treated firms as control firms for late treated ones. Therefore, our estimates are not impacted by 2×2 DID comparisons of previously treated groups to newly treated groups that typically confound staggered DID applications (Goodman-Bacon (2021)).

TABLE 4
Difference-in-Differences: The Effect on Short-Selling Activity

Table 4 provides evidence that the informativeness of short-selling demand increased after the introduction of satellite data. We report coefficient estimates based on the DID regression model in equation (2). The dependent variable $EARET_{itq}$ is the cumulative factor-adjusted return from 1 trading day before to 1 trading week after the quarterly earnings announcement. $POST_{itq}$ is an indicator variable that takes the value 1 after the initiation of satellite coverage, $TREAT_{itq}$ is an indicator variable that takes the value 1 for the treated group of retailers with satellite coverage, $\Delta SHORT_{itq}^{Dem}$ is the cumulative change in the lender quantity on loan from the end of the quarter to 2 days before the quarterly announcement, and $\Delta SHORT_{itq}^{Sup}$ is the cumulative change in the available supply of lendable share from the end of the quarter to 2 days before the quarterly announcement. We report regression results using the standardized z-values of the continuous predictors. For each retailer in the treated group, we use a symmetric event window before and after the initiation of satellite coverage. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. The matched control group includes an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit GICS industry and do not have satellite coverage. We obtain information about named competitors from FactSet Revere. We report t-statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, based on 2-tailed tests.

$Z_{itq} =$	Dependent Variable: $EARET_{itq}$	
	$\Delta SHORT_{itq}^{Dem}$	$\Delta SHORT_{itq}^{Sup}$
Z_{itq}	-0.005** (-2.44)	-0.005 (-1.53)
$POST_{itq} \times Z_{itq}$	0.001 (0.13)	0.000 (0.06)
$TREAT_{itq} \times Z_{itq}$	0.011** (2.42)	-0.009 (-1.13)
$POST_{itq} \times TREAT_{itq} \times Z_{itq}$	-0.024** (-2.67)	0.008 (0.68)
Characteristics (C_{itq})	Yes	Yes
Firm fixed effects (θ_i)	Yes	Yes
Quarter fixed effects (δ_q)	Yes	Yes
Adj. R^2	6.1%	5.8%
No. of obs.	4,139	4,152

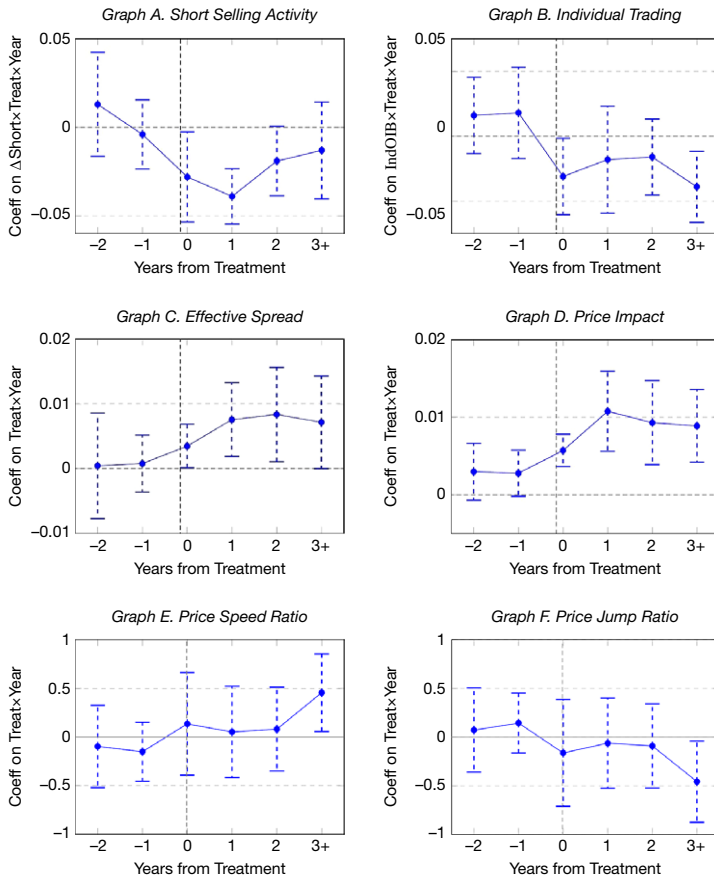
control groups (e.g., Roberts and Whited (2013)).⁷ While we cannot formally test for the parallel-trends assumption, we can evaluate whether pre-treatment trends in outcomes of interest are the same for covered and control firms. Graph A of Figure 5 presents the estimated treatment effects on short-selling activity in the years surrounding the introduction of satellite data. The pre-treatment coefficients are indistinguishable from 0, which implies that covered and control retailers are indistinguishable from each other prior to the introduction of satellite data. Turning to the post-treatment coefficients, we observe a sharp shift in the informativeness of short-selling activity that persists in the years after the release of satellite data.

In summary, the DID tests so far provide evidence that the ability of short sellers to profit from targeting retailers with bad news for the quarter increased after the introduction of satellite data for the group of retailers with satellite coverage. The parallel-trends analysis suggests that the observed effects are more likely to be a result of the introduction of satellite data as opposed to an alternative force. Next, we attempt to identify the effect of satellite data on individual investor trading.

⁷More broadly, the DID design does not represent a magic bullet that solves the selection problem (Roberts and Whited (2013)). In our setting, the selection problem comes from the fact that data vendors choose which firms to cover and when to begin selling satellite imagery data. As a result, there is no exogenous source of variation in satellite coverage such that firms are randomly assigned as treated or control. Another challenge is that the matched-control firms may be covered by different forms of alternative data. For example, credit card transaction data might be available for firms that do not have satellite coverage.

FIGURE 5
Parallel-Trends Analysis

Figure 5 presents the estimated treatment effects and corresponding confidence intervals across key outcomes, including the impact of the introduction of satellite data on the effective spread and price impact, the informativeness of short selling and individual trading activity, as well as the price speed and price jump ratios. We plot the estimated treatment effects and corresponding 95% confidence intervals in the years surrounding the release of satellite data.



E. Identifying the Effect on Individual Investor Trading

The substantial costs of acquiring, processing, and integrating satellite data mean that such data can be seen as a form of private information. These costs make it especially difficult for individual investors with limited resources to access the satellite data, making them less informed than sophisticated investors who can afford to incur the substantial costs of acquiring, processing, and integrating such data. The notion of unequal access to satellite data relates to Kyle's (1985) model of the dynamics surrounding insider trading. This model predicts that noise traders will camouflage the trades of insiders who will then profit at the noise traders' expense. In our setting, the satellite data takes the role of inside information for the sophisticated investors with access, and the individual investors without access operate as noise traders. It follows that the introduction of satellite data could lead to individual investor trades becoming less informative.

To estimate the effect of satellite coverage on the informativeness of individual trading activity, we use the following DID regression model:

(3)
$$\begin{aligned} \text{EARET}_{iq} = & \alpha + \beta_1 \text{POST}_{iq} + \beta_2 \text{TREAT}_{iq} + \beta_3 \text{POST}_{iq} \times \text{TREAT}_{iq} + \\ & \beta_4 \text{IndOIB}_{iq} + \beta_5 \text{POST}_{iq} \times \text{IndOIB}_{iq} + \beta_6 \text{TREAT}_{iq} \times \text{IndOIB}_{iq} + \\ & \beta_7 \text{POST}_{iq} \times \text{TREAT}_{iq} \times \text{IndOIB}_{iq} + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}. \end{aligned}$$

The independent variable IndOIB_{iq} measures abnormal individual order imbalance as the average individual order imbalance over the pre-earnings announcement window from the end of the quarter to 2 days before the earnings announcement adjusted for the average individual order imbalance during the quarter. We use the BJZZ algorithm to identify individual trades and measure the daily individual order imbalance as individual buys minus individual sells divided by the total number of shares outstanding.

The coefficient on the triple interaction $\text{POST}_{iq} \times \text{TREAT}_{iq} \times \text{IndOIB}_{iq}$ captures the change in the informativeness of individual trading activity across the treated and matched control groups after the introduction of satellite coverage. A decrease in the informativeness of individual trading activity in the stock of retail companies with satellite coverage would imply that individual trading is more negatively related to earnings announcement returns in the post period (i.e., $\beta_7 < 0$).

Table 5 reports the DID regression results for equation (3). Column 1 reports that the coefficient on the triple interaction $\text{POST}_{iq} \times \text{TREAT}_{iq} \times \text{IndOIB}_{iq}$ is

TABLE 5
Difference-in-Differences: The Effect on Individual Trading

Table 5 provides evidence that the informativeness of individual trading decreased after the introduction of satellite data. We report coefficient estimates based on the DID regression model in equation (3). The dependent variable EARET_{iq} is the cumulative factor-adjusted return from 1 trading day before to 1 trading week after the quarterly earnings announcement. POST_{iq} is an indicator variable that takes the value 1 after the initiation of satellite coverage, TREAT_{iq} is an indicator variable that takes the value 1 for the treated group of retailers with satellite coverage, IndOIB_{iq} is the abnormal individual order imbalance, IndBUYS_{iq} is the abnormal individual buying activity, and IndSELLS_{iq} is the abnormal individual selling activity over the pre-earnings announcement window from the end of the quarter to 2 days before the earnings announcement. We use Boehmer's et al. (2021) method to identify the individual order flow in the TAQ data. We report regression results using the standardized z-values of the continuous predictors. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. We report regression results using the standardized z-values of the continuous predictors. For each retailer in the treated group, we use a symmetric event window before and after the initiation of satellite coverage. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. The matched control group includes an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit SIC industry and do not have satellite coverage. We obtain information about named competitors from FactSet Revere. We report t-statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, based on 2-tailed tests.

$Z_{iq} =$	Dependent Variable		
	IndOIB_{iq}	IndBUYS_{iq}	IndSELLS_{iq}
Z_{iq}	-0.015*** (-3.64)	-0.013** (-2.44)	-0.003 (-1.14)
$\text{POST}_{iq} \times Z_{iq}$	0.015** (2.94)	0.013* (2.01)	0.005 (0.79)
$\text{TREAT}_{iq} \times Z_{iq}$	0.019** (2.95)	0.019*** (3.51)	0.009 (1.30)
$\text{POST}_{iq} \times \text{TREAT}_{iq} \times Z_{iq}$	-0.026*** (-3.39)	-0.024*** (-3.76)	-0.013 (-1.73)
Characteristics (C_{iq})	Yes	Yes	Yes
Firm fixed effects (θ_i)	Yes	Yes	Yes
Quarter fixed effects (δ_q)	Yes	Yes	Yes
Adj. R^2	6.7%	6.5%	6.1%
No. of obs.	3,877	3,877	3,877

significantly negative, which is consistent with a decrease in the informativeness of individual trading for the treated group of retailers with satellite coverage after the introduction of satellite data. In contrast, the significantly positive coefficient on the interaction $\text{POST}_{it} \times \text{IndOIB}_{it}$ implies that the informativeness of individual trading has actually improved over time for the control group of retailers with no satellite coverage. Put differently, while there is a general positive trend in the informativeness of individual trading activity (e.g., Kelley and Tetlock (2013)), this trend is actually offset for retailers with satellite coverage.

Focusing on the change in the differential predictive ability of individual trading activity, a 1-standard-deviation increase in individual trading is associated with a 2.6% lower earnings announcement return for treated retailers relative to the matched control group after the introduction of satellite data. In absolute terms, this magnitude is 6.2 times the value of the average earnings announcement return. Figure 1 further shows that there is no evidence of significant pre-treatment effects and that the informativeness of individual trading drops with the introduction of satellite data.

To probe the dynamics of individual trading, we separately examine the effect of the introduction of satellite data on the informativeness of individual buying activity (IndBUYS_{it}) versus individual selling activity (IndSELLS_{it}). Columns 2 and 3 of Table 5 provide evidence that the overall decrease in the informativeness of individual trading is primarily due to changes in the informativeness of individual buying activity. The coefficient on the triple interaction term is significantly negative for IndBUYS_{it} and it is indistinguishable from 0 for IndSELLS_{it} .

We hasten to note that while the BJZZ algorithm is the most popular tool available to identify individual trades in TAQ, recent studies raise issues with the effectiveness of the algorithm to identify individual trades (e.g., Battalio, Jennings, Saglam, and Wu (2022), Barber, Huang, Jorion, Odean, and Schwarz (2023), and Barardehi, Bernhardt, Da, and Warachka (2023)).⁸ As Barber et al. (2023) point out, any study on individual trading activity is a joint test of two null hypotheses. Adopted to our setting, the joint test is that i) individual trading activity is the same for stocks with and without satellite coverage before and after the introduction of satellite data, and ii) the ability of the BJZZ algorithm to identify individual trading is the same for stocks with and without satellite coverage. An important assumption for the validity of our identification is that the performance of the BJZZ algorithm, in terms of false negative and false positive classification errors, does not systematically differ across treated and control firms. Without access to the true individual trading, we cannot test the differential performance of the BJZZ algorithm.

⁸Barber et al. (2023) ran a trading experiment from Dec. 2021 to June 2022, and show that the BJZZ algorithm is prone to false negative identification with 65% of individual trades not classified as such. For the month of Dec. 2020, Battalio et al. (2022) use proprietary data from order flow wholesalers and provide evidence that the BJZZ can also generate false positive errors by classifying nonretail as retail activity. In addition, Barardehi et al. (2023) provide evidence that the BJZZ algorithm differentially identifies a subset of individual orders that order flow wholesalers internalize to provide liquidity to institutions.

Notwithstanding this limitation, we build on Barber et al. (2023) and provide a series of tests that help alleviate concerns about the validity of our inferences. These tests focus on individual buying activity, which is the primary driver of changes in the informativeness of individual order. For our first test, following their recommendation, we repeat our tests using the quote midpoint method of Lee and Ready (1991) to sign trades instead of the sub-penny digit method of BJZZ. Barber et al. (2023) also observe that the BJZZ algorithm generates higher misclassification rates for firms with spreads wider than 10 cents. Based on this observation, we repeat our tests focusing firms with spreads below 10 cents and 5 cents. Finally, Barber et al. (2023) note that the performance of the BJZZ algorithm is compromised for stocks included in the SEC's Tick Size Pilot (TSP). Accordingly, we also repeat our tests after excluding TSP stocks. Table A5 in the Supplementary Material reports these sensitivity tests and shows consistent evidence of a decrease in the informativeness of individual buying activity following the introduction of satellite data.

Viewed as a whole, the evidence is consistent with a decrease in the informativeness of individual trading activity following the introduction of satellite data. Next, we examine the effect of the introduction of satellite data on stock liquidity.⁹

F. Identifying the Effect on Stock Liquidity

So far, our DID tests provide evidence that the dynamics of short-selling activity and individual trading activity closely mirror each other. Whereas the ability of short sellers to anticipate bad news for retailers with satellite coverage has improved after the introduction of satellite data, individual investors' trades have become less informative with respect to anticipating retailer news for the quarter. The opposing impact of the introduction of satellite data on short-selling activity versus individual trading activity implies that differential access to alternative data can lead to an increase in information asymmetry among market participants.

Prior research concludes that greater information asymmetry among market participants leads to lower stock liquidity (e.g., Copeland and Galai (1983), Glosten and Milgrom (1985), Kyle (1985), and Easley and O'Hara (1987)). In these models, information asymmetry can increase in either the proportion of informed traders or the precision of their information. Given that the information asymmetry between sophisticated and unsophisticated investors is likely to increase before the earnings announcement (Lee, Mucklow, and Ready (1993)), we expect a decrease in stock liquidity for treated firms in the trading days leading to earnings announcements. We use the following DID regression model to estimate the effect of satellite coverage on stock liquidity:

⁹To be clear, we do not argue that individual investors are the only market participants on the other side of informed short sellers. Rather, we argue that due to the expensive acquisition and processing costs, individual investors are less likely to have access to big data. Thus, our evidence does not preclude that unequal access to big data may also increase information asymmetry among groups of sophisticated investors.

$$(4) \quad \text{LIQ}_{iq} = \alpha + \beta_1 \text{POST}_{iq} + \beta_2 \text{TREAT}_{iq} + \beta_3 \text{POST}_{iq} \times \text{TREAT}_{iq} \\ + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}.$$

We estimate [equation \(4\)](#) using two complementary measures of stock liquidity. First, we compute the effective spread of a firm's trades using intraday trading data from the TAQ database. We measure the effective spread as the daily firm average of $2 \times (|P_k - M_k|)$, where P_k the price on the observed trade and M_k is the midpoint of the National Best Bid and Offer (NBBO) quotes for that trade (SPREAD_{iq}). Relative to the quoted bid–ask spread, the effective spread offers a more accurate measure of stock liquidity since trades are often executed within the quoted spread (e.g., Petersen and Fialkowski (1994)). We then compute the average effective spread from the day after the end of the quarter to 2 days before the earnings announcement. This pre-announcement measurement window is the same used for the short selling and individual investor order flow tests and allows us to examine the window over which sophisticated investors are likely to have the greatest information advantage. Second, we examine whether the price impact on a trade, the permanent component of the effective spread, is affected by the introduction of satellite data. We follow Holden and Jacobsen (2014) and measure price impact (PRICE_IMPACT_{iq}) as the daily average of $2 \times D_k(M_{k+5} - M_k)$, where M_{k+5} is the midpoint 5 minutes after M_k , and D_k is equal to +1 if we identify the trade as buyer initiated and –1 if it is seller initiated.

[Table 6](#) reports the DID regression results for [equation \(4\)](#). Starting with column 1, the significantly positive coefficient on the interaction $\text{POST}_{iq} \times \text{TREAT}_{iq}$ is consistent with an increase in the effective spread and, therefore, a decrease in stock liquidity for the treated group of retailers with satellite coverage after the introduction of satellite data. The inclusion of time-fixed effects

TABLE 6
Difference-in-Differences: The Effect on Stock Liquidity

[Table 6](#) provides evidence that stock liquidity decreased after the introduction of satellite data. We report coefficient estimates based on the DID regression model in [equation \(4\)](#). The dependent variable in column 1 is the effective spread, measured as 2 times the absolute difference between the price on a trade and the midpoint of the National Best Bid and Offer (NBBO) quotes for that trade. The dependent variable in column 2 is price impact, which measures the permanent component of the effective spread by comparing the midpoint 5 minutes after the trade to the midpoint at the time of the trade. Both measures are computed from the day after the end of the quarter to 2 days before the earnings announcement. POST_{iq} is an indicator variable that takes the value 1 after the initiation of satellite coverage, TREAT_{iq} is an indicator variable that takes the value 1 for the treated group of retailers with satellite coverage. For each retailer in the treated group, we use a symmetric event window before and after the initiation of satellite coverage. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. The matched control group includes an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit GICS industry and do not have satellite coverage. We obtain information about named competitors from FactSet Revere. We report *t*-statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, based on 2-tailed tests.

	Dependent Variable	
	SPREAD_{iq}	PRICE_IMPACT_{iq}
$\text{POST}_{iq} \times \text{TREAT}_{iq}$	0.584%*** (3.19)	0.667%*** (3.89)
Characteristics (C_{iq})	Yes	Yes
Firm fixed effects (θ_i)	Yes	Yes
Quarter fixed effects (δ_q)	Yes	Yes
Adj. R^2	79.7%	74.8%
No. of obs.	3,726	3,726

alleviates concerns that our result is due to an aggregate time-trend in the effective spread. Turning to column 2, we find consistent evidence of an increase in price impact and, therefore, a decrease in stock liquidity for the treated group of retailers with satellite coverage after the introduction of satellite data. Given an average effective spread (price impact) of 0.031 (0.023), the increase in effective spread (price impact) after the introduction of satellite data represents a shift of 19% (29%) from the unconditional mean. Graphs C and D of Figure 5 confirm that there is no evidence of significant pre-treatment effects, and that stock liquidity drops with the introduction of satellite data.

The effect of satellite coverage on information asymmetry is likely to vary with the information environment of each retailer. In particular, we expect that the introduction of satellite data provides sophisticated investors with a greater edge when uncertainty about firm fundamentals is higher. We use firm size, age, and volatility to proxy for fundamental uncertainty. We conjecture that the effect of satellite data on information asymmetry will be more pronounced for smaller, younger, and high-volatility retailers for which fundamental uncertainty is likely to be higher. To test this conjecture, we expand equation (4) as follows:

(5)
$$LIQ_{iq} = \alpha + \beta_1 POST_{iq} + \beta_2 TREAT_{iq} + \beta_3 POST_{iq} \times TREAT_{iq} + \beta_4 F_{iq} + \beta_5 POST_{iq} \times F_{iq} + \beta_6 TREAT_{iq} \times F_{iq} + \beta_7 POST_{iq} \times TREAT_{iq} \times F_{iq} + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}.$$

TABLE 7
Difference-in-Differences: Heterogeneous Effect on Stock Liquidity

Table 7 provides evidence that the effect of satellite data on stock liquidity is more pronounced for smaller, younger, and high-volatility retailers. We report coefficient estimates based on the DID regression model in equation (5). We use two measures of stock liquidity: the effective spread (SPREAD_{iq}) and price impact (PRICE_IMPACT_{iq}). We measure the effective spread as 2 times the absolute difference between the price on a trade and the midpoint of the NBBO quotes for that trade. We measure price impact as the permanent component of the effective spread by comparing the midpoint 5 minutes after the trade to the midpoint at the time of the trade. Both measures are computed from the day after the end of the quarter to 2 days before the earnings announcement. POST_{iq} is an indicator variable that takes the value 1 after the initiation of satellite coverage, TREAT_{iq} is an indicator variable that takes the value 1 for the treated group of retailers with satellite coverage. F_{iq} denotes our indicators of a retailer's information environment. Focusing on the treated group of retailers, the I(SMALL_{iq}) indicator identifies retailers with below-median market capitalization, the I(YOUNG_{iq}) indicator identifies retailers with below-median firm age, and the I(HVLT_{iq}) indicator identifies retailers with above-median stock return volatility. We measure age relative to the firm's founding year using data available from Jay Ritter's website. We measure volatility as the standard deviation of the past 12 months prior to the quarter. For each retailer in the treated group, we use a symmetric event window before and after the initiation of satellite coverage. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. The matched control group includes an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit GICS industry and do not have satellite coverage. We obtain information about named competitors from FactSet Revere. We report *t*-statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, based on 2-tailed tests.

$F_{iq} =$	Dependent Variable: SPREAD _{iq}			Dependent Variable = PRICE_IMPACT _{iq}		
	I(SMALL _{iq})	I(YOUNG _{iq})	I(HVLT _{iq})	I(SMALL _{iq})	I(YOUNG _{iq})	I(HVLT _{iq})
POST _{iq} × TREAT _{iq}	0.200% (0.91)	−0.072% (−0.32)	0.234% (0.86)	0.375%* (1.82)	0.199% (1.09)	0.416%** (2.20)
POST _{iq} × TREAT _{iq} × F _{iq}	0.798%*** (4.42)	1.691%*** (4.04)	0.712%** (2.27)	0.607%*** (4.70)	1.207%*** (3.79)	0.536%** (2.75)
Characteristics (C _{iq})	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effects (θ _i)	Yes	Yes	Yes	Yes	Yes	Yes
Quarter fixed effects (δ _q)	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R ²	79.9%	80.2%	80.7%	74.9%	75.3%	74.9%
No. of obs.	3,726	3,726	3,717	3,726	3,726	3,717

Equation (5) interacts the baseline model in equation (4) with F_{iq} , which denotes the alternative indicators of a retailer's information environment. Focusing on the treated group, $I(\text{SMALL}_{iq})$ identifies retailers with below-median market capitalization, $I(\text{YOUNG}_{iq})$ identifies retailers with below-median firm age, and $I(\text{HVLTV}_{iq})$ identifies retailers with above-median stock return volatility. For the dependent variable, LIQUIDITY_{iq} , we continue to use the effective spread and price impact as our measures of stock liquidity. A significantly positive value for β_5 would suggest that the introduction of satellite data led to a larger increase in information asymmetry for firms with higher fundamental uncertainty.

Table 7 presents the cross-sectional tests and provides evidence of heterogeneous effects on stock liquidity. The estimated coefficient on the triple interaction term $\text{POST}_{iq} \times \text{TREAT}_{iq} \times F_{iq}$ is significantly positive for all three indicators of fundamental uncertainty. The evidence is consistent with the conjecture that the effect of satellite data on information asymmetry is concentrated in smaller, younger, and high-volatility retailers for which fundamental uncertainty is likely to be higher. These results are consistent with the idea that the introduction of satellite data led to a larger increase in information asymmetry for firms with higher fundamental uncertainty.

To summarize, the stock liquidity results provide evidence consistent with a significant increase in information asymmetry around the quarterly reports of retailers with satellite coverage, after the introduction of such data. These effects are concentrated in firms with higher fundamental uncertainty. We hasten to note that we do not interpret our results as de facto evidence of a one-to-one transfer of wealth from individual investors to short sellers. Rather, we argue that due to the expensive acquisition and processing costs, individual investors are less likely to have access to big data. More broadly, our evidence does not preclude that unequal access to big data may also increase information asymmetry among groups of sophisticated investors.

G. Identifying the Effect on Price Discovery

Timelier price discovery would imply that stock prices impound more information prior to the earnings announcement, so that the reaction to the public earnings announcement itself is muted. This argument is consistent with rational expectations models showing that the price reaction to public announcements is decreasing in the amount of pre-announcement information (e.g., Kim and Verrecchia (1991)).

We consider two complementary price discovery measures based on Weller (2018). Anchoring on the earnings announcement day (day 0), we measure the PRICE_SPEED_RATIO as the cumulative abnormal return between days $t - 10$ and $t - 2$ divided by the cumulative abnormal return between days $t - 10$ and $t + 1$. We measure the PRICE_JUMP_RATIO as the cumulative abnormal return between days $t - 1$ and $t + 1$ relative to day 0 divided by the cumulative abnormal return between days $t - 10$ and $t + 1$. To mitigate the impact of small values in the denominator of the ratios, we require absolute values of cumulative returns between days $t - 10$ and $t + 1$ in excess of 0.10%. An increase in the speed of price discovery would imply higher values for the PRICE_SPEED_RATIO and lower values for

TABLE 8
Difference-in-Differences: The Effect on the Speed of Price Discovery

Table 8 provides evidence that the introduction of satellite data had no detectable effect on the speed of price leading to the quarterly earnings announcements of retailers with satellite coverage. We report coefficient estimates based on the DID regression model in equation (6). We measure the PRICE_SPEED_RATIO as the cumulative factor-adjusted return between trading days $t - 10$ and $t - 2$ relative to the earnings announcement (day 0) divided by the cumulative factor-adjusted return between trading days $t - 10$ and $t + 1$. We measure the PRICE_JUMP_RATIO as the cumulative factor-adjusted return between trading days $t - 1$ and $t + 1$ relative to day 0 divided by the cumulative factor-adjusted return between trading days $t - 10$ and $t + 1$. To mitigate small denominator problems, we require absolute values in excess of 0.10% for the denominator in the price speed and price jump ratios. $POST_{iq}$ is an indicator variable that takes the value 1 after the initiation of satellite coverage, $TREAT_{iq}$ is an indicator variable that takes the value 1 for the treated group of retailers with satellite coverage. The treated group of retailers includes 650 firm-quarter observations from 2011:Q1 to 2017:Q4. The matched control group includes an average of 2.4 size-matched competitors per retailer that operate in the same 6-digit GICS industry and do not have satellite coverage. We obtain information about named competitors from FactSet Revere. Column 1 reports results using the true $POST_{iq}$ indicator. We report t -statistics in parentheses based on clustered standard errors by time. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively, based on 2-tailed tests.

	Dependent Variable	
	PRICE_SPEED_RATIO _{iq}	PRICE_JUMP_RATIO _{iq}
$POST_{iq} \times TREAT_{iq}$	0.205 (1.20)	-0.210 (-1.19)
Characteristic controls	Yes	Yes
Firm fixed effects	Yes	Yes
Quarter fixed effects	Yes	Yes
Adj. R^2	2.8%	2.7%
No. of obs.	4,259	4,259

the PRICE_JUMP_RATIO, since a greater fraction of information would be incorporated in the pre-announcement window.

To estimate the effect of the introduction of satellite coverage on price discovery, we use the following model:

(6)
$$Y_{iq} = \alpha + \beta_1 POST_{iq} + \beta_2 TREAT_{iq} + \beta_3 POST_{iq} \times TREAT_{iq} + C_{iq} + \theta_i + \delta_q + \varepsilon_{iq}.$$

Table 8 reports the DID regression results for equation (6). Starting with the PRICE_SPEED_RATIO, column 1 reports that the coefficient on the interaction $POST_{iq} \times TREAT_{iq}$ is indistinguishable from 0. This finding is consistent with our earlier evidence of limited pre-announcement price activity with the majority of trading taking place on earnings announcements (Graph B of Figure 4). Turning to the PRICE_JUMP_RATIO, column 2 reports consistent evidence that the coefficient on the interaction $POST_{iq} \times TREAT_{iq}$ is also indistinguishable from 0.

Overall, we find that on average the introduction of satellite data had no detectable effect on the speed of price discovery leading to the quarterly reports of retailers with satellite coverage and the amount of information impounded in stock returns prior to the public disclosure of retailer performance. This null result, however, is not definitive evidence that no change happened since failure to reject the null does not constitute accepting it. In fact, it is plausible that the null result masks a gradual, long-term impact on price discovery that gets attenuated when averaging across years in the post-treatment period.

To evaluate this possibility, we turn to Graphs E and F of Figure 5 and evaluate the estimated treatment effects in the years surrounding the introduction of satellite data. The evidence shows that price discovery does not improve immediately, but

gradually over time. In particular, we find evidence of a significant increase in the price speed ratio accompanied by a significant decrease in the price jump ratio 3 years after the introduction of satellite data. The evidence suggests that the introduction of satellite data did not immediately enhance stock price discovery, but rather its effect on price informativeness unfolded over time with the increased adoption of alternative data.

IV. Conclusion

We study the introduction of satellite coverage of major U.S. retailers as a source of big data in capital markets. Our evidence shows that satellite data enabled sophisticated investors to formulate profitable trading strategies, especially by targeting the quarterly reports of retailers with bad news for the quarter. Using a DID design, we find that the introduction of the satellite data led to more informed short-selling activity, less informed individual buying activity, and lower stock liquidity around the reports of retailers with satellite coverage. Overall, our article provides evidence that unequal access to big data can increase information asymmetry among market participants without immediately enhancing price discovery. Our evidence adds to research on the role of big data in capital markets and contributes to the ongoing debate on the impact of technology and data abundance on capital markets.¹⁰

Over time, the value of publicized signals should decay (McLean and Pontiff (2016)). Still, data hunters will continue to scour data from anywhere there is a digital footprint and sophisticated investors will continue to invest in data in their quest to gain an edge. As the market's most sophisticated players come to rely on sources of data that are ever more out of reach for the general public, regulators and policymakers may need to grapple with the question of what the social welfare implications of unequal access to big data are.

Supplementary Material

To view supplementary material for this article, please visit <http://doi.org/10.1017/S0022109023001448>.

References

- Ahn, B. H.; R. M. Bushman; and P. N. Patatoukas. "Under the Hood of Activist Fraud Campaigns: Private Information Quality, Disclosure Incentives and Stock Lending Dynamics." Working Paper, U.C. Berkeley (2023).
- Ahn, B. H., and P. N. Patatoukas. "Identifying the Effect of Stock Indexing: Impetus or Impediment to Arbitrage and Price Discovery?" *Journal of Financial and Quantitative Analysis*, 57 (2022), 2022–2062.

¹⁰Whereas our article does not explore social welfare implications, the consequences of big data may extend beyond the capital markets. Mihet (2022) proposes a theory whereby innovations in financial technology can exacerbate capital income inequality between more sophisticated investors, who can afford to acquire costly private information, and less sophisticated investors, who have little access to private information.

- Banerjee, S.; J. Davis; and N. Gondhi. "When Transparency Improves, Must Prices Reflect Fundamentals Better?" *Review of Financial Studies*, 31 (2018), 2377–2414.
- Barardehi, Y. H.; D. Bernhardt; Z. Da; and M. Warachka. "Uncovering the Liquidity Premium in Stock Returns Using Retail Liquidity Provision." Available at SSRN 4057713 (2023).
- Barber, B. M.; X. Huang; P. Jorion; T. Odean; and C. Schwarz. "A (Sub) Penny for Your Thoughts: Tracking Retail Investor Activity in TAQ." Available at SSRN 4202874 (2023).
- Battalio, R.; R. Jennings; M. Saglam; and J. Wu. "Identifying Market Maker Trades As "Retail" From Taq: No Shortage of False Negatives and False Positives." Working Paper, University of Notre Dame (2022).
- Boehmer, E.; C. Jones; X. Zhang; and X. Zhang. "Tracking Retail Investor Activity." *Journal of Finance*, 75 (2021), 2249–2305.
- Cao, S.; W. Jiang; J. L. Wang; and B. Yang. "From Man vs. Machine to Man+Machine: The Art and AI of Stock Analyses." NBER Working Paper No. w28800 (2021).
- Chang, R., and Z. Da. "Nowcasting Firms' Fundamentals: Evidence from the Cloud." Working Paper, University of Notre Dame (2022).
- Copeland, T. E., and D. Galai. "Information Effects on the Bid–Ask Spread." *Journal of Finance*, 38 (1983), 1457–1469.
- Dessaint, O.; T. Foucault; and L. Frésard. "Does Alternative Data Improve Financial Forecasting? The Horizon Effect." Available at SSRN 3702411 (2021).
- Dugast, J., and T. Foucault. "Data Abundance and Asset Price Informativeness." *Journal of Financial Economics*, 130 (2018), 367–391.
- Dugast, J., and T. Foucault. "Equilibrium Data Mining and Data Abundance." HEC Paris Research Paper No. FIN-2020-1393, Université Paris-Dauphine Research Paper (3710495) (2021).
- Easley, D., and M. O'Hara. "Price, Trade Size, and Information in Securities Markets." *Journal of Financial Economics*, 19 (1987), 69–90.
- Froot, K.; N. Kang; G. Ozik; and R. Sadka. "What Do Measures of Real-Time Corporate Sales Say about Earnings Surprises and Post-Announcement Returns?" *Journal of Financial Economics* 125 (2017), 143–162.
- Gerken, W. C., and M. Painter. "The Value of Differing Points of View: Evidence from Financial Analysts' Geographic Diversity." Available at SSRN 3479352 (2019).
- Glosten, L. R., and P. R. Milgrom. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics*, 14 (1985), 71–100.
- Gofman, M.; G. Segal; and Y. Wu. "Production Networks and Stock Returns: The Role of Vertical Creative Destruction." *Review of Financial Studies*, 33 (2020), 5856–5905.
- Goodman-Bacon, A. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, 225 (2021), 254–277.
- Grennan, J., and R. Michaely. "FinTechs and the Market for Financial Analysis." *Journal of Financial and Quantitative Analysis*, 56 (2021), 1877–1907.
- Grossman, B. S. J., and J. E. Stiglitz. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70 (1980), 393–408.
- Hayek, F. A. "The Use of Knowledge in Society." *American Economic Review*, 35 (1945), 519–530.
- Holden, C. W., and S. Jacobsen. "Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions." *Journal of Finance*, 69 (2014), 1747–1785.
- IHS Markit. "Demystifying Alternative Data: Can Alternative Data Really Enhance Your Investment Strategy?" Available at the IHS Markit Website (2019).
- Jank, S.; C. Roling; and E. Smajlbegovic. "Flying Under the Radar: The Effects of Short-Sale Disclosure Rules on Investor Behavior and Stock Prices." *Journal of Financial Economics*, 139 (2021), 209–233.
- Jones, C. M.; A. V. Reed, and W. Waller. "Revealing Shorts: An Examination of Large Short Position Disclosures." *Review of Financial Studies* 29 (12), 3278–3332 (2016).
- Kang, J. K.; L. Stice-Lawrence; and Y. T. F. Wong. "The Firm Next Door: Using Satellite Images to Study Local Information Advantage." *Journal of Accounting Research*, 59 (2021), 713–750.
- Kelley, E. K., and P. C. Tetlock. "How Wise Are Crowds? Insights from Retail Orders and Stock Returns." *Journal of Finance*, 68 (2013), 1229–1265.
- Kim, O., and R. E. Verrecchia. "Trading Volume and Price Reactions to Public Announcements." *Journal of Accounting Research*, 29 (1991), 302–321.
- Kyle, A. S. "Continuous Auctions and Insider Trading." *Econometrica: Journal of the Econometric Society*, 53 (1985), 1315–1335.
- Lee, C. M., and Ready, M. J. "Inferring trade direction from intraday data". *The Journal of Finance*, 46 (2), 733–746 (1991).
- Lee, C. M.; B. Mucklow; and M. J. Ready. "Spreads, Depths, and the Impact of Earnings Information: An Intraday Analysis." *Review of Financial Studies*, 6 (1993), 345–374.

- McLean, R. D., and J. Pontiff. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, 71 (2016), 5–32.
- Mihet, R. "Financial Technology and the Inequality Gap." Working Paper. Available at SSRN 3474720 (2022).
- Mukherjee, A.; G. Panayotov; and J. Shon. "Eye in the Sky: Private Satellites and Government Macro Data." *Journal of Financial Economics*, 141 (2021), 234–254.
- Pedersen, L. H., *Efficiently Inefficient*. Princeton, NJ: Princeton University Press (2015).
- Petersen, M. A., and D. Fialkowski. "Posted Versus Effective Spreads: Good Prices or Bad Quotes?" *Journal of Financial Economics*, 35 (1994), 269–292.
- Roberts, M. R., and T. M. Whited. "Endogeneity in Empirical Corporate Finance." In *Handbook of the Economics of Finance*, Vol. 2, G. Constantinides, M. Harris, and R. Stulz, eds. Amsterdam, The Netherlands: Elsevier (2013), 493–572.
- Verrecchia, R. E. "Information Acquisition in a Noisy Rational Expectations Economy." *Econometrica* 50 (1982), 1415–1430.
- Walton, S., and J. Huey, *Made in America*, Vol. 216. New York: Doubleday (1992).
- Weller, B. M., "Does Algorithmic Trading Reduce Information Acquisition?" *Review of Financial Studies*, 31 (2018), 2184–2226.
- Zhu, C. "Big Data as a Governance Mechanism." *Review of Financial Studies*, 32 (2019), 2021–2061.