Alternative Data, Firms' Ownership and Stock Price Fragility

Maxime Bonelli and Thierry Foucault*

August 7, 2026

Preliminary Draft

Please Do Not Circulate or Cite Without the Authors' Permission

ABSTRACT

We study how technological innovations in information production affect firms' ownership structures. Focusing on the expansion of commercial alternative data, we develop a novel text-based measure of firm-level alternative data coverage. We show that greater alternative data coverage reallocates ownership away from long-term investors and toward hedge funds, which trade more actively and shorten their holding horizons. While this ownership reallocation reduces stock return comovement with the market and industry, consistent with greater incorporation of firm-specific information into prices, it also increases stock price fragility: During market downturns, hedge funds reduce their holdings more aggressively, making highly covered firms' stock prices more sensitive to market-wide shocks. Overall, our findings show that ownership reallocation is one channel through which changes in firms' information environment can increase stock price fragility.

Keywords: Big data, ownership, hedge funds, market fragility.

JEL classification: G11, G14, G23, G32

*Bonelli is affiliated with London Business School and can be reached at mbonelli@london.edu. Foucault is affiliated with HEC Paris can be reached at foucault@hec.fr. Bonelli thanks the Data Science and AI Initiative, the Institute of Entrepreneurship and Private Capital, and the Research and Materials Development Grant at London Business School for financial support. We thank Abel Gonzales-Hishinuma and Hector Choné-Ducasse for excellent research assistance. All rights reserved by Maxime Bonelli and Thierry Foucault.

I Introduction

Recent years have witnessed major technological innovations in information production, driven by the combination of increasingly abundant data, powerful algorithms for extracting signals from these data, and advances in computing power. While these innovations have fundamentally transformed how information is produced and processed, an important question is how they affect capital markets (see Goldstein et al., 2021). In this paper, we show that they reshape firms' ownership structures by attracting data-savvy institutional investors and, in turn, alter the sensitivity of firms' stock prices to market-wide shocks.

We focus on one manifestation of these technological innovations: the rapid expansion of commercial alternative data, which enables investors to obtain information about firms from sources that lie largely outside firms' control. Unlike traditional sources of information, such as regulatory filings, accounting data, and corporate disclosures, alternative data provide higher frequency information on firms' activities from external digital sources, including social media, online product reviews, point-of-sale transactions, and satellite imagery. We develop a novel text-based measure of firms' coverage by commercial alternative datasets and use it to study how this new informational environment reshapes firms' ownership structures and stock price dynamics.

To build this measure, we rely on data from Datarade (https://datarade.ai/), one of the largest digital marketplaces for alternative datasets.[1] As of November 2024, Datarade hosted approximately 4,000 datasets offered by more than 500 data providers. We compare the textual description of each commercially available dataset with the business description reported in each U.S. firm's 10-K filings submitted to the Securities and Exchange Commission (SEC), using text embeddings. We then aggregate the similarities across all datasets available at a given point in time to construct a firm-quarter measure of alternative data coverage spanning the period from 1994 to 2023.

As expected, alternative data coverage increases over time, reflecting the growing availability of alternative datasets. More importantly, cross-sectional heterogeneity also increases substantially. By 2023, the difference in coverage between the 95th and 5th

[1] Liu et al. (2025) also use data from Datarade, but to study the market for commercial data providers and their pricing strategies.

percentiles reaches approximately six standard deviations, compared with less than one standard deviation in 2005. There is also considerable variation across industries. Firms in the Finance, Energy, and Utilities sectors are among the most covered, whereas firms in Consumer Nondurables and Manufacturing are among the least covered on a relative basis.

We then use this measure to study how a firm's coverage by alternative data affects its ownership by institutional investors. Using alternative data effectively requires specialized expertise in data science as well as substantial upfront investments in data infrastructure, including storage, computing capacity, and access to large data libraries.[2] We therefore expect data-savvy investors, such as hedge funds and quantitative asset managers, to be more likely to acquire alternative datasets and use them when investing in firms covered by these data.[3] Moreover, because alternative data should enable these investors to form more accurate forecasts of future returns, we expect them to take larger long or short positions, depending on the realization of their signals.[4] Consequently, we expect data-savvy investors to hold a larger fraction of the shares of firms with greater alternative data coverage.

To test this hypothesis, we use data on U.S. institutional equity holdings from the FactSet Ownership database and classify institutional investors into five categories, following Koijen et al. (2024): investment advisers, hedge funds, long-term investors (pension funds and insurance companies), private banks, and brokers. We define a firm as "treated" if, in a given quarter, it becomes "highly covered", that is, enters the highest tercile of the alternative data coverage distribution within its industry. We then examine whether data-savvy investors (hedge funds) increase the weight of treated firms in their portfolios following the treatment, relative to other firms in their investment universe and relative to other institutional investors, by controlling for firm-quarter and investor-quarter fixed effects.

Consistent with our hypothesis, we find a significant increase in hedge funds' holdings of treated firms after the treatment. Moreover, the effect is even stronger for hedge funds that

[2]Consistent with this view, a survey by the Alternative Investment Management Association ("Casting the Net: How Hedge Funds are Using Alternative Data") notes that "*Working with alternative data in an efficient manner (i.e. in a way to produce either investment or operational alpha) requires having in place the appropriate infrastructure (talent, as well as financial resources) [...]*".

[3]Hedge funds are major users of alternative data. See "Casting the Net: How Hedge Funds are Using Alternative Data" and, for instance, "Hedge funds scour alternative data for edge on Covid and economy", Financial Times, August 4, 2020. Goldstein et al. (2021) also note that big data provides more information for sophisticated players.

[4]See, for instance, Dugast and Foucault (2025) for a model generating this prediction.

are more likely to rely on quantitative investment strategies.[5] This suggests that investors with a comparative advantage in acquiring and processing alternative data respond more strongly to increases in firms' alternative data coverage.

The previous finding is consistent with the idea that access to alternative data enables hedge funds to form more precise signals about future stock returns and, consequently, to take more aggressive long positions when their signals predict positive abnormal returns. As a result, hedge funds hold, on average, larger positions in treated firms, since short positions are not reflected in ownership data. Because alternative data generate both positive and negative information about firms, greater signal precision and frequency should induce hedge funds to adjust their positions more frequently once a firm becomes treated.[6] Consistent with this prediction, we find that when a firm becomes treated, hedge funds' turnover in the stock increases by more than 2 percentage points (p.p.), while the probability that they initiate or liquidate a position rises by 0.6 p.p, relative to other firms in their investment universe and relative to other institutional investors. They also shorten their average holding period by approximately two quarters (16% relative to the sample average).

The previous tests examine whether alternative data coverage induces heterogeneous portfolio responses across institutional investors. However, they do not directly show how alternative data coverage affects the composition of a firm's ownership. To address this question, we examine how high coverage affects the ownership shares of the different types of institutional investor at the firm-quarter level. Specifically, we compare each treated firm with a set of control firms matched on market capitalization and pre-treatment ownership composition using a stacked difference-in-differences design.

Consistent with the holdings-level evidence, we find that, post treatment, treated firms experience a significant increase of about 0.37 percentage points (p.p.) in hedge funds' ownership share, relative to control firms (13% relative the sample median). In contrast, the ownership shares of most other investor categories decline, although these declines are not statistically significant. In additional analyses, we find a statistically significant decrease in

[5]To identify hedge funds that are more likely to follow a quantitative investment approach, we use the text of investment advisers' regulatory Form ADV Part 2A filings, following Chen et al. (2026).

[6]Indeed, hedge funds are unlikely to adjust their positions when signals are weak because of trading costs. As signal precision increases, signals are therefore more likely to trigger a buy or sell decision rather than a no-trade decision.

the ownership share of investors that have held the firm's stock for at least eight quarters, i.e., long-term shareholders. Combined with our earlier finding that hedge funds shorten their holding periods following a firm's treatment, this result suggests that greater alternative data coverage increases firms' exposure to short-term shareholders. We next examine the implications of this ownership reallocation for covered firms' stock price dynamics.

As greater alternative data coverage increases the role of hedge funds in a firm's ownership, it should reduce return comovement if alternative data enhance hedge funds' ability to produce firm-specific information.[7] If this interpretation is correct, the reduction in return comovement should weaken when financing constraints (Shleifer and Vishny, 1997; Brunnermeier and Pedersen, 2009) or internal risk limits (Gârleanu and Pedersen, 2007) force hedge funds to reduce their positions, typically during market downturns.[8] Moreover, because greater alternative data coverage increases hedge fund ownership, it may also amplify firms' exposure to hedge fund fire sales during these episodes, making their stock prices more fragile.[9]

We find support for these implications. Specifically, post treatment, treated firms exhibit significantly lower return comovement with the market, their industry peers, and firms with similar alternative data coverage, relative to control firms, consistent with greater incorporation of firm-specific information into prices. This finding is consistent with the hypothesis that alternative data enable sophisticated investors, such as hedge funds, to acquire more precise private information about firms' idiosyncratic fundamentals.

However, this effect is concentrated when market returns are positive. During periods of poor market performance—defined as quarters in the bottom decile of the distribution of market returns—the reduction in return comovement disappears. This pattern suggests that hedge funds reduce the production of firm-specific information during market downturns. Consistent with hedge funds withdrawing capital during these episodes, we find that hedge funds become significantly larger net sellers of treated firms during periods of poor market

[7] Many alternative datasets are primarily useful for predicting firm-specific rather than market-wide outcomes. For example, Bonelli and Foucault (2025) show that satellite imagery data on retailers have no predictive power for aggregate market returns.

[8] See Boyson et al. (2010) and Kruttli et al. (2025) for evidence that financing constraints and internal risk limits lead hedge funds to deleverage.

[9] Greenwood and Thesmar (2011) define a fragile asset as "*susceptible to non-fundamental shifts in demand*," for instance because its owners are exposed to correlated shocks. As noted by Ben-David et al. (2021), this implies larger price declines during periods of market stress.

performance following treatment, compared to control firms. As a result, treated firms experience significantly larger declines in stock prices during market downturns post treatment. Therefore, treated stocks exhibit greater sensitivity to negative market returns. Together, these findings suggest that alternative data coverage generate a trade-off: by attracting sophisticated investors, they improve the incorporation of firm-specific information into prices during normal times, while simultaneously making treated firms more exposed to hedge fund fire sales during periods of market downturns.

Related Literature Our paper first contributes to the literature on the effects of alternative data on the finance industry. Past studies have examined how this new source of information impacts analysts' forecasts (Cao et al., 2024; Chi et al., 2024; Dessaint et al., 2024), price informativeness (Gao and Huang, 2019; Zhu, 2019; Katona et al., 2024), asset managers' performance (Bhojraj et al., 2012; Bonelli and Foucault, 2025), and venture capitalists' investment decisions (Bonelli, 2026).[10] However, very few studies have analyzed how the expansion of alternative data affects the firms being covered. Zhu (2019) documents that alternative data have a disciplining effect on corporate managers. To our knowledge, our paper is first to use alternative data coverage as a laboratory to understand how innovations in information technologies affect firms' ownership and market fragility.

A growing literature examines how firms' accumulation of internal data affects their performance and valuation (Begenau et al., 2018; Cong et al., 2021; Farboodi and Veldkamp, 2026; Mihet et al., 2026; Eisfeldt et al., 2026). Our focus is different. Rather than studying the consequences of firms' own data assets, we examine how a permanent external change in firms' informational environment (coverage by alternative datasets) affects their ownership structure and, ultimately, the sensitivity of their stock prices to market downturns.[11]

This paper also contributes to the literature on institutional ownership. A large body of work examines how institutional investors shape firms' policies and governance (e.g., Bushee,

[10] In addition, many papers study the predictive power for future cash-flows or stock returns of different types of alternative data. See, for instance, Katona et al. (2025) for satellite imagery, Green et al. (2019) for online employer reviews, or Huang (2018) for customers' online reviews.

[11] Our paper is also related to Farboodi et al. (2022), who develop a structural measure of investors' information. Our paper identifies one channel through which the information environment changes—the expansion of commercial alternative data—and studies how these innovations alter firms' ownership structures and, through this ownership channel, affect stock price dynamics and market fragility.

1998; Aghion et al., 2013; Chen et al., 2020; Liu et al., 2020), while another strand studies the role of specific investor types, including activist hedge funds (Brav et al., 2008; Becht et al., 2009; Klein and Zur, 2009; Becht et al., 2017) and short-horizon investors (Cella et al., 2013; Derrien et al., 2013; Cremers et al., 2020). Our paper makes two contributions to this literature. First, we identify a novel determinant of firms' ownership structure by showing that increases in commercial alternative data coverage reallocate ownership away from long-term investors and toward hedge funds. Second, we document the consequences of this ownership reallocation. Our findings complement those of Cella et al. (2013), who show that firms with more short-term investors perform worse during periods of market stress. We show that this ownership structure also has important benefits during normal times: by attracting sophisticated investors who trade on firm-specific information, greater alternative data coverage improves the incorporation of information into stock prices, thereby reducing return comovement.[12]

Last, our paper contributes to the literature on stock price fragility (Brunnermeier and Pedersen, 2009; Greenwood and Thesmar, 2011; Ben-David et al., 2021). Existing studies (Cella et al., 2013; Ben-David et al., 2021; Glossner et al., 2025) show that institutional ownership can be a source of stock price fragility because institutional investors are exposed to correlated liquidity shocks and may engage in synchronized sales during periods of market stress. Our contribution is complementary but distinct. Rather than taking ownership structure as given and focusing on the proximate cause of fragility, we identify one of its underlying determinants. Specifically, we show that greater coverage by alternative data modifies firms' ownership structures by attracting hedge funds, thereby increasing these firms' exposure to hedge fund fire sales during market downturns.

The rest of the paper is organized as follows. Section II describes the data used in the paper. Section III develops our text-based measure of firm-level alternative data coverage. Section IV examines how institutional investors adjust their portfolio holdings and trading behavior following increases in firms' alternative data coverage. Section V studies how these portfolio adjustments reshape firms' ownership structures. Section VI investigates the implications of these ownership changes for stock return comovement and stock price fragility. Section VII concludes.

[12]Informative stock prices can benefit firms in several ways. For instance, they facilitate the design of more efficient managerial compensation contracts (see Dow et al., 2024).

II Data

A Commercial Alternative Data Marketplace

Our primary source of information on commercial alternative datasets is Datarade (https://datarade.ai), one of the largest online marketplaces for alternative data. At the time of our data collection (November 11th, 2024), the platform listed close to 4,000 datasets supplied by over 500 distinct data providers across a broad range of industries and applications. We choose Datarade for three reasons. First, it provides substantially broader coverage than existing directories of commercial alternative data. Internet Appendix A compares Datarade with several competing platforms. Second, Datarade reports rich metadata for most datasets, including provider identity, historical coverage, categories, intended use cases, and detailed textual descriptions. Third, these textual descriptions are sufficiently long and detailed to allow us to systematically match datasets to publicly listed firms using modern text-embedding techniques.

We develop a web crawler to systematically collect all publicly available information from the Datarade platform. Specifically, we download the webpage associated with every dataset listed on the marketplace together with its metadata. The resulting database contains 3,860 unique datasets and includes information on the dataset name, provider, category classification, intended use cases, historical coverage, and both short and long textual descriptions. These textual descriptions summarize the type of information contained in each dataset and constitute the primary input for our measure of firm-level alternative data coverage described in Section III.

The datasets span a wide range of update frequencies, ranging from real-time streams to annual releases. Internet Appendix Table A.2 reports the distribution of update frequencies. More than one quarter of all datasets are updated at an intraday frequency (real time, secondly, or minutely), while more than 40% are updated at least daily, illustrating the high-frequency nature of much of the commercial alternative data.

In addition to dataset-level information, Datarade maintains dedicated webpages for each data provider. We similarly collect all provider pages, obtaining information on the 540 distinct providers offering the datasets in our sample. These pages contain provider

descriptions, data offerings, and, in many cases, information on the provider's founding year. We use this information to document the evolution of the commercial alternative data industry over time.

To measure the evolution of the commercial alternative data industry over time, we identify the date at which each data provider first became active. We combine three complementary sources of information. First, we extract founding dates directly from providers' textual descriptions available on Datarade. We use a large language model to identify references to company founding, establishment, or incorporation dates.[13] Second, for each provider, we identify its official homepage and retrieve the corresponding internet domain registration date from publicly available records. Domain registration dates provide an objective measure of when providers established their commercial online presence.[14] Finally, for the small number of providers for which neither approach yields a start date, we manually identify the provider's founding year using publicly available information. Overall, this procedure yields provider start dates for more than 90% of all providers listed on Datarade.[15]

Figure I presents the evolution of the commercial alternative data industry over time. Panel A reports the number of new data providers entering the market each year, while Panel B reports the cumulative number of active providers. The figure documents a remarkable expansion of the industry over the past two decades. Fewer than 100 providers were active before 2000, whereas approximately 500 providers were listed on Datarade in 2020. This rapid growth illustrates the increase in the supply of commercially available alternative data over time.

Because providers may release multiple datasets sequentially over time, provider entry alone does not determine when a specific dataset becomes commercially available. We

[13]Specifically, we leverage the GPT-4o model via the OpenAI API to process the text descriptions and extract founding, establishment, or incorporation dates for each provider.

[14]For each provider, we first identify its official homepage and extract the corresponding domain name. We then query the publicly available WHOIS database through the https://who.is service to retrieve the domain registration date (reported under fields such as "Creation Date," "Registered On," or "Domain Name Commencement Date").

[15]Whenever both the textual descriptions and the domain registration approach identify a start date, we compare the two measures. Differences primarily arise because many providers were incorporated before registering their current internet domain. Since domain registration more closely captures the date at which providers became commercially active online, we use the domain registration date, when available, as our preferred start date measure.

therefore construct a dataset-specific release year. A dataset is considered commercially available in year t if (i) its provider was active before year t, and (ii) the dataset's reported historical coverage begins before year t.

B Firm Business Descriptions

To characterize firms' economic activities, we construct a comprehensive database of annual business descriptions from firms' Form 10-K filings. We retrieve all 10-K filings submitted to the U.S. Securities and Exchange Commission (SEC) through the EDGAR filing system between 1994 and 2023. For each filing, we extract Item 1 ("Business"), which provides a standardized description of the firm's principal products, services, markets, and business activities. We focus on Item 1 because it is specifically intended to describe the firm's operations and business model (see Hoberg and Phillips, 2016) and therefore offers the most appropriate textual representation for matching firms to commercially available alternative datasets.

Extracting Item 1 from SEC filings is non-trivial because filings are submitted in heterogeneous HTML and text formats, and the structure of individual filings varies considerably across firms and over time. We therefore develop an automated parsing algorithm that identifies the beginning and end of the Item 1 section using the standardized section headers required by SEC filings while accounting for the numerous formatting conventions observed in practice. The parser further removes HTML tags, tables, exhibits, cross-references, and other filing-specific markup to recover a clean textual description of each firm's business activities.

We then apply a series of standard text preprocessing procedures. Specifically, we convert all text to lowercase, remove punctuation, numbers, HTML entities, and non-alphabetic characters, eliminate common English stopwords, and normalize whitespace. We retain the remaining words without imposing any finance-specific dictionary or manually selected keywords.

The final corpus contains one business description for each firm-year observation and constitutes the firm-side input used to construct our measure of alternative data coverage described in Section III. Because firms update their Item 1 disclosures annually to reflect

changes in their operations, acquisitions, product offerings, and strategic focus, the resulting textual representation evolves over time.

C Institutional Holdings

We obtain data on U.S. institutional equity holdings from the FactSet Ownership database, which consolidates Securities and Exchange Commission (SEC) Form 13F filings to provide security-level holdings for institutional investors. Form 13F requires institutional investment managers with at least $100 million in assets under management to report their quarterly equity holdings.[16]

Following prior work our sample begins in 2000-Q1. We follow Koijen et al. (2024) to classify institutional investors into five broad categories: investment advisors (including mutual funds), hedge funds, long-term investors (pension funds and insurance companies), private banks, and brokers.[17]

In our empirical analysis at the investor–firm–quarter level, we follow Koijen and Yogo (2019) and define an investor's investment universe in each quarter as the set of stocks that the investor currently holds or has held at any point during the previous eleven quarters. Consequently, investor–firm observations corresponding to stocks that belong to the investment universe but are not currently held are retained in the sample with a portfolio weight equal to zero. Including these zero holdings avoids conditioning the analysis on positive positions and allows us to distinguish between intensive-margin adjustments and complete exits.

For each investor–firm–quarter observation, we compute the market value of the position, portfolio weights, excess portfolio weights relative to market capitalization-based weight, trading activity, entry and exit decisions, and holding duration. Aggregating these holdings

[16] Following Koijen et al. (2024), we exclude two FactSet entity identifiers (0FSVG4-E and 000V4B-E) that contain known reporting errors relative to the original EDGAR filings.

[17] The classification follows the FactSet investor subtype codes used by Koijen et al. (2024). Investment advisors include subtype codes IC, RE, PP, SB, and MF; hedge funds include AR, FH, FF, FU, and FS; long-term investors include FO, SV, and IN; private banks include CP, FY, and VC; and brokers include BM, IB, ST, and MM. Private banking further includes private banking and wealth management, family offices, private equity, and venture capital. In our analyses at the holding level, we exclude observations for investors with a missing FactSet investor subtype code, as these institutions cannot be reliably assigned to one of the five investor categories.

across investors allows us to construct firm-level measures of ownership composition, including ownership shares by investor type and shareholder turnover.

Panel A of Table I reports summary statistics for the investor–stock–quarter panel. The average portfolio weight is 0.35%, while the average excess portfolio weight relative to market capitalization is close to zero by construction. Conditional on belonging to an investor's investment universe, positions are held for an average of 12.8 quarters. Investment advisors constitute the largest investor group, accounting for 62% of investor–firm–quarter observations, followed by private banks (15%), hedge funds (10%), brokers (8%), and long-term investors (6%).

We complement the ownership data with firm-level accounting and market information from the FactSet database. We use these data to construct the firm characteristics used in our empirical analyses, including market capitalization, shares outstanding, market stock price, leverage (debt-to-assets), dividend payments, and book equity. Panel B of Table I reports summary statistics at the stock–quarter level.

III Measuring Firm-Level Alternative Data Coverage

A Methodology

Our objective is to construct a time-varying measure of the extent to which publicly listed firms are covered by commercially available alternative datasets. Ideally, one would directly observe, for every dataset, the universe of firms to which it applies and the date at which each firm start being covered by the dataset. Such information, however, is not publicly disclosed. Commercial data providers typically describe the contents and intended applications of their datasets through textual descriptions, but they do not explicitly report the firms covered by each dataset.

We therefore propose a text-based measure of firm-level alternative data coverage. The key intuition is that datasets describing information closely related to a firm's business activities are more likely to contain economically relevant information about that firm. For example, a dataset described as providing "satellite imagery of retail parking lots" is naturally more relevant for a retailer than for a pharmaceutical company, while a dataset

covering "commercial aircraft flight data" is more relevant for airlines than for utilities. By systematically quantifying the semantic similarity between firms' business descriptions and commercial dataset descriptions, we infer the extent to which each firm is covered by the expanding universe of commercially available alternative data.

Recent advances in natural language processing make it possible to represent entire documents as vectors in a low-dimensional semantic space, where distances reflect semantic similarity rather than simple word overlap. Unlike traditional bag-of-words approaches, text embeddings capture semantic rather than lexical similarity. For example, a firm's 10-K may describe its activities using the terms "grocery stores" or "supermarkets," while a commercial dataset may instead refer to "food retail." Text embedding methods recognize that these descriptions refer to closely related economic activities, whereas traditional measures based solely on word overlap would substantially underestimate their similarity.

Following the text-based finance literature (e.g., Hoberg and Phillips, 2025), we adopt this embedding approach to compare firms' business descriptions with commercial alternative datasets. Specifically, we train a 300-dimensional Doc2Vec model (see for instance Le and Mikolov, 2014) on the corpus of dataset descriptions collected from Datarade.[18] We then use the trained model to infer embedding vectors for firms' Item 1 business descriptions, thereby representing both firms and datasets in the same semantic vector space. Documents describing similar economic activities are therefore represented by nearby vectors, even when they use different vocabularies or expressions.

Let $\mathbf{f}_{it} \in \mathbb{R}^{300}$ denote the embedding associated with firm i's business description in year t, and let $\mathbf{d}_j \in \mathbb{R}^{300}$ denote the embedding associated with dataset j. We measure the relevance of dataset j for firm i using the cosine similarity between the corresponding embedding vectors:

$$Similarity_{ijt} = \frac{\mathbf{f}_{it}' \mathbf{d}_j}{\|\mathbf{f}_{it}\| \, \|\mathbf{d}_j\|}. \tag{1}$$

$Similarity_{ijt}$ takes values between 0 and 1, with larger values indicating that the dataset description is more closely related to the firm's business activities and is therefore more likely to contain information relevant for valuing the firm. Internet Appendix B provides examples

[18] We choose 300 dimensions following Hoberg and Phillips (2025).

of high and low similarities between datasets and firms' descriptions.

Because the universe of commercially available datasets expands and firms' activities evolve over time, firm-level coverage must also evolve over time. Let $\mathcal{D}_t$ denote the set of datasets that are commercially available in year t, where commercial availability is defined according to the procedure described in Section II.A. Specifically, a dataset belongs to $\mathcal{D}_t$ if its provider was active before year t and the dataset's historical coverage begins before year t.

We define firm i's alternative data coverage in year t as

$$Coverage_{it} = \sum_{j \in \mathcal{D}_t} Similarity_{ijt}. \tag{2}$$

This measure captures two distinct dimensions of alternative data availability. First, coverage increases as new commercially available datasets enter the marketplace. Second, coverage is higher when those datasets are more closely related to the firm's underlying business activities. Consequently, firms operating in market segments for which many commercially available datasets exist receive higher coverage scores than firms for which little alternative data are available.

B The Evolution of Alternative Data Coverage

B.1 Distribution over Time and across Industries

The coverage measure defined in equation (2) captures the cumulative semantic similarity between a firm's business activities and the universe of commercially available alternative datasets. By construction, the measure is expected to increase over time as new datasets become commercially available and therefore enter the sum in equation (2). At the same time, we expect firms operating in sectors for which alternative data are particularly informative should exhibit systematically higher coverage than other firms.

To examine these properties, Panel A of Figure II reports the evolution of the distribution of the raw coverage measure over time. To facilitate interpretation, we standardize the raw coverage measure over the full sample to have zero mean and unit variance. Consequently, a

one-unit increase corresponds to a one-standard-deviation increase in a firm's exposure to commercially available alternative data relative to the full sample.

Panel A of Figure II reveals two important patterns. First, the distribution of alternative data coverage shifts steadily upward over time, reflecting the rapid expansion of the commercial alternative data industry documented in Figure I. This finding confirms that firms become increasingly exposed to commercially available datasets as the data market expands.

Second, the cross-sectional dispersion of coverage also increases substantially over time. While the median firm's exposure rises gradually, firms in the upper tail experience much larger increases in coverage than those in the lower tail. By the end of the sample period, the gap between the 95th and 5th percentiles is about six standard deviations. This growing dispersion suggests that the expansion of commercial alternative data has been highly uneven across firms, creating substantial heterogeneity in firms' information environments.

Panel B of Figure II reports the distribution of the (standardized) coverage measure across Fama–French 12 industries. Industries are ranked by their median coverage. Several differences emerge. Financial firms exhibit the highest levels of alternative data coverage, followed by firms in the energy sector, whereas non-durables and manufacturing display lower median coverage. These patterns suggest that industries such as finance and energy generate rich digital footprints—including payment flows, financial transactions, energy production, shipping activity, and satellite imagery—that can be commercialized by alternative data providers. By contrast, firms in more traditional manufacturing industries might generate fewer such digital data. The figure also reveals substantial within-industry heterogeneity. Particularly among industries with high median coverage, firms differ clearly in their exposure to commercially available datasets, reflecting differences in business models and the types of alternative data they relate to.

B.2 Determinants of Alternative Data Coverage

After documenting the time-series and cross-sectional patterns in Figure II, we next examine whether our coverage measure correlates with observable firm characteristics. A possibility is that firms with greater alternative data coverage are larger or systematically different along other observable dimensions. For example, larger firms may generate richer digital footprints

and therefore attract greater attention from commercial data providers. Similarly, firms with different asset structures may differ systematically in the availability of related alternative datasets.

To examine this possibility, Table II regresses the coverage measure (standardized over the full sample) on a set of firm characteristics. Columns (1)–(4) report cross-sectional regressions including year-quarter fixed effects, while Columns (5)–(8) additionally include firm fixed effects, thereby identifying the coefficients from within-firm changes over time.

The cross-sectional results show that larger firms tend to exhibit greater alternative data coverage. In the univariate specifications, market capitalization and total assets are positively associated with coverage. This is also the case for book-to-market. However, these correlations change once the three characteristics are included jointly. In Column (4), assets remain strongly positively associated with coverage, while market capitalization and book-to-market becomes negatively associated. Overall, the cross-sectional results indicate that alternative data coverage is correlated with observable firm characteristics, although these variables explain only a modest fraction of the cross-sectional variation (R^2 of 0.16 in Column (4)).

When firm fixed effects are introduced, changes in assets and book-to-market are still positively correlated with alternative data coverage but to a lower extent. In the multivariate specification reported in Column (8), the estimated coefficients are substantially smaller than in the corresponding cross-sectional regression in Column (4). Changes in market capitalization and total assets are no longer statistically significant. Overall, these results suggest that although alternative data coverage is correlated with observable firm characteristics in the cross section, the within-firm variation is only weakly associated with changes in standard firm fundamentals.

C Constructing the Treatment Variable

The previous subsection shows that the raw coverage measure captures three distinct sources of variation. First, coverage increases mechanically over time as the commercial alternative data industry expands and new datasets become available. Second, some industries are systematically more exposed to alternative data because their activities naturally generate

richer digital footprints. Third, even within industries, firms differ substantially in their exposure to commercially available datasets. Our empirical strategy is designed to isolate this third source of variation.

Accordingly, we standardize the raw coverage measure within each Fama–French 48 industry and calendar year-quarter. Specifically, for every industry–year-quarter pair, firms are ranked according to their raw coverage measure and assigned to terciles of the resulting distribution. This transformation removes the secular increase in alternative data coverage over time as well as persistent differences in coverage across industries, allowing firms to be compared only with their direct industry peers in a given quarter.

Our baseline treatment variable exploits firms experiencing relatively high alternative data coverage. Specifically, a firm is classified as "covered" whenever it belongs to the highest tercile of the coverage distribution within its Fama–French 48 industry and calendar year-quarter. The treatment indicator, $Covered \times Post$, equals one whenever the firm belongs to the highest tercile and zero otherwise. We consider that firms exit treatment if their relative exposure to available alternative data is no longer in the top highest tercile.

In practice, treatment status is highly persistent. The quarter-to-quarter transition matrix shows that firms in the highest coverage tercile have a 94.7% probability of remaining in the same tercile in the following quarter. Persistence is equally high for firms in the lowest tercile. Only about 1% of firms move directly between the lowest and highest terciles from one quarter to the next, while transitions between the second to the third tercile occur with a probability of roughly 4%. Therefore, changes in treatment status are infrequent. They reflect shifts in firms' relative exposure to commercially available alternative data, typically following the introduction of new datasets that are particularly relevant to a firm's business activities. Internet Appendix C provides examples of key datasets triggering firms' treatment.

This treatment definition has two advantages. First, it abstracts from the aggregate growth of the commercial alternative data industry, ensuring that our estimates are not driven by the secular increase in data availability documented in Figure I. Second, by comparing firms only with other firms operating in the same industry and quarter, it removes persistent differences in alternative data intensity across industries. Consequently, our empirical analyses exploit changes in a firm's relative exposure to alternative data rather than differences arising from industry composition or aggregate time trends.

Finally, Panel B of Table I reports that the treatment indicator equals one for 16.7% of firm–quarter observations. This fraction is lower than one third because firms differ in the number of quarters they are present in the sample.

IV Alternative Data and Institutional Portfolio Holdings

A Average Effect on Institutional Portfolio Holdings

Our first objective is to examine how institutional investors adjust their portfolios following increases in firms' alternative data coverage. We begin by estimating a difference-in-differences specification that compares institutional portfolio allocations before and after firms become highly covered by commercial alternative datasets:

$$Y_{i,j,t} = \beta \left\{ Covered \times Post \right\}_{j,t} + \Gamma' X_{j,t} + \alpha_j + \gamma_{i \times t} + \varepsilon_{i,j,t}, \tag{3}$$

where $Y_{i,j,t}$ denotes the portfolio weight allocated by investor i to stock j in year-quarter t, or the excess portfolio weight relative to the stock's market weight. The indicator $Covered \times Post$ equals one after stock j enters the highest tercile of alternative data coverage. The vector $X_{j,t}$ includes control variables for firm characteristics, namely log market capitalization, market beta, leverage (debt-to-assets ratio), dividend to book equity (the ratio of annual dividends per split-adjusted share times shares outstanding to book equity), log book equity, and lagged one-quarter stock returns. Standard errors are three-way clustered by investor, stock, and year-quarter.

The specification includes stock fixed effects, α_j, which absorb all time-invariant stock characteristics, and investor×year-quarter fixed effects, $\gamma_{i \times t}$, which absorb all time-varying investor-level shocks, including changes in assets under management, investor flows, and portfolio-wide reallocations. Consequently, identification comes from comparing how a given investor reallocates capital across covered and non-covered stocks within the same quarter.

Table III reports the results from the baseline difference-in-differences specification. Across all institutional investors, alternative data coverage is associated with a small decline in

portfolio weights. The estimated effect is marginally significant (10% level) once control variables are included (Column (2)), and becomes statistically insignificant once portfolio weights are expressed relative to the stock's market weight. These findings suggest that, on average, institutional investors do not systematically increase their exposure to firms experiencing greater alternative data coverage.

This aggregate result, however, may mask substantial heterogeneity across investor types. If alternative data primarily benefit sophisticated investors capable of acquiring and processing these new information sources, increases in holdings by such investors may be offset by reductions in holdings by other institutions. In addition, although this specification provides a useful benchmark, its interpretation remains limited because the timing of alternative data coverage may be endogenous to time-varying firm characteristics. For example, firms may become covered as they experience changes in their economic fundamentals that simultaneously affect institutional demand. While the inclusion of firm characteristics as controls alleviates some of these concerns, it cannot fully account for unobserved time-varying shocks at the firm level. As a result, the estimated treatment effect may reflect both the impact of alternative data coverage and other (non-observed) contemporaneous changes in firms' characteristics. To address these concerns, our main empirical strategy, described below, exploits differences across investor types within the same stock and quarter.

B Heterogeneity Across Investor Types

The absence of an average effect is not inconsistent with alternative data affecting institutional ownership. Rather, it may reflect heterogeneous responses across different types of investors. Commercial alternative datasets are costly to acquire and require substantial investments in data infrastructure, engineering, and statistical or machine learning expertise before they can be transformed into investment signals. Consequently, data-savvy investors are likely to derive greater benefits from these datasets than investors relying on more traditional information sources. Bonelli and Foucault (2025) show that this leads naturally to an equilibrium in which data-savvy investors invest more in alternative data than traditional investors. Consistent with this prediction, they find that traditional investors experience a significant decline in performance on stocks once they become covered by alternative data, whereas this effect is

much weaker for data-savvy investors.

We expect hedge funds to be particularly data-savvy because they invest heavily in proprietary data, quantitative research, and computational infrastructure. As the availability of commercial alternative datasets expands, hedge funds should therefore increase their portfolio allocations toward firms for which these datasets are most informative, while other investors may respond less strongly or even reduce their holdings. Our central hypothesis is that the expansion of alternative data reallocates ownership toward investors with a comparative advantage in processing these new information sources. This motivates our preferred empirical specification, which exploits heterogeneity in portfolio responses across investor types within the same stock and quarter.

B.1 Empirical Strategy

Specifically, we ask whether investors with a comparative advantage in acquiring and processing alternative data—such as hedge funds—increase their holdings of newly covered firms more than other institutional investors. Our empirical strategy therefore exploits heterogeneity across investor types within the same stock and quarter. Rather than comparing covered and uncovered firms, it compares how different investor types adjust their holdings of the same stock following an increase in alternative data coverage. Identification thus comes from the differential portfolio response of data-savvy investors relative to other institutional investors facing the same investment opportunity. We estimate the following triple-difference specification:

$$Y_{i,j,t} = \beta \left\{ Covered \times Post \right\}_{j,t} \times HedgeFund_i + \alpha_{j \times t} + \gamma_{i \times t} + \varepsilon_{i,j,t}, \tag{4}$$

where $Y_{i,j,t}$ denotes the portfolio weight allocated by investor i to stock j in quarter t, or the excess portfolio weight relative to the stock's market weight. $HedgeFund_i$ is an indicator equal to one if investor i is classified as a hedge fund (see Section II.C). Standard errors are three-way clustered by investor, stock, and quarter.

As in specification (3), the regression includes investor×quarter fixed effects, $\gamma_{i \times t}$, which absorb all time-varying investor-level shocks, including changes in assets under management,

investor flows, and other portfolio-wide reallocations. Importantly, specification (4) additionally includes stock×quarter fixed effects, $\alpha_{j\times t}$, which absorb all observed and unobserved time-varying stock characteristics, including economic fundamentals, news, and the timing of alternative data coverage itself. Consequently, the interaction term $\{Covered \times Post\}_{j,t}$ is fully absorbed by these fixed effects, so identification no longer relies on comparing covered and uncovered firms. Instead, it comes from comparing how different investor types adjust their holdings of the same stock in the same quarter, relative to their own holdings of other stocks in their investment universe. The identifying assumption is therefore substantially weaker than in the baseline difference-in-differences specification, as all contemporaneous stock-level shocks are absorbed by the stock×quarter fixed effects.

The coefficient of interest, β, captures whether hedge funds increase their holdings of newly covered firms more than other institutional investors. We focus on hedge funds because they are generally viewed as the institutional investors best equipped to acquire, process, and exploit commercial alternative data. Relative to other investor types, hedge funds make greater use of quantitative investment strategies, invest more heavily in data infrastructure, and are typically more intensive users of non-traditional information sources. If alternative data generate economically valuable information, we therefore expect hedge funds to trade covered firms' stocks more than other institutional investors.

B.2 Empirical Results

Table IV reports the results. The evidence strongly supports the hypothesis that alternative data coverage differentially affects institutional portfolio allocation across investor types. Relative to other investors, hedge funds significantly increase their holdings of firms experiencing high alternative data coverage. The estimated increase in portfolio weights is approximately 3 basis points and is highly statistically significant (Column (1)). An almost identical effect is obtained when portfolio weights are measured relative to the stock's market weight (Column (3)), indicating that the results are not driven by changes in firms' market capitalization or benchmark weights.[19]

Columns (2) and (4) allow the effect to vary across all investor categories, using investment

[19]We note that this result is expected because our specification includes stock×quarter fixed effects which control for any stock-level time-varying characteristics, including market, index or benchmark weight.

advisors as the omitted category. Relative to investment advisors, hedge funds increase their portfolio weights in covered firms by approximately 3 basis points. In contrast, we find no evidence that long-term investors increase their exposure to covered firms. If anything, brokers and private banks reduce their portfolio weights following increases in alternative data coverage, with economically and statistically significant declines of 0.9 and 1.9 basis points, respectively. These findings indicate that the absence of an average treatment effect documented in Table III masks substantial heterogeneity across investor types.

Overall, the evidence suggests that alternative data coverage does not uniformly increase institutional demand. Instead, it reallocates ownership toward hedge funds at the expense of other institutional investors. This pattern is consistent with the view that commercially available alternative datasets primarily benefit investors with a comparative advantage in acquiring, processing, and exploiting these new information sources.

B.3 Dynamics of the Effect

To examine the dynamics of hedge fund portfolio adjustments, we replace the post-treatment indicator with a series of event-time indicators corresponding to the quarters before/after a firm first enters the highest tercile of alternative data coverage. We estimate the following event-study specification:

$$Y_{i,j,t} = \sum_{k=-8, k\neq -1}^{12+} \beta_k \left\{ Covered \times Quarter(k) \right\}_{j,t} \times HedgeFund_i + \alpha_{j\times t} + \gamma_{i\times t} + \varepsilon_{i,j,t}, \quad (5)$$

where $Quarter(k)$ denotes event time relative to the quarter in which stock j first enters the highest tercile of alternative data coverage. The omitted category is the quarter immediately preceding treatment ($k = -1$). Standard errors are three-way clustered by investor, stock, and quarter.

Because all time-varying stock characteristics are absorbed by stock×quarter fixed effects, the coefficients β_k are identified by comparing how hedge funds and other institutional investors adjust their holdings of the same stock within the same quarter before/after coverage, relative to their other portfolio holdings.

Figure III presents the estimated event-study coefficients β_k using the stock's excess portfolio weight relative to its market weight as the dependent variable. Two findings emerge. First, we find no evidence of differential pre-trends. The coefficients during the eight quarters preceding treatment are economically small and statistically indistinguishable from zero, providing support for the assumption that hedge funds do not systematically increase their exposure to future treated firms prior to the expansion of alternative data coverage.

Second, hedge fund portfolio allocations increase gradually following treatment. The estimated coefficients become positive immediately after coverage and continue to increase over subsequent quarters, reaching approximately four basis points after three years. This pattern is consistent with firms progressively entering the investment universe of hedge funds as commercially available alternative datasets become relevant to their business activities.

Importantly, the observed increase in hedge fund portfolio weights should not be interpreted as evidence that hedge funds simply purchase and hold these stocks. Alternative datasets may generate both positive and negative signals, implying that hedge funds may either increase or reduce their positions depending on the information revealed by the data. Because our holdings data only observe long positions, the ownership response captures only one dimension of hedge fund trading activity. We therefore next examine whether alternative data coverage also affects hedge funds' trading behavior, focusing on portfolio turnover, entry and exit decisions, and holding duration.

C Effect on Institutional Investors' Trading Behavior

The previous results establish that hedge funds increase their portfolio weights in firms experiencing high alternative data coverage. We next investigate whether this increase reflects a more active trading strategy. If alternative datasets provide investors with more frequent, granular, and firm-specific information, one would expect hedge funds not only to hold larger positions in covered firms but also to adjust these positions more actively over time. This prediction is particularly relevant because many commercial alternative datasets are updated at high frequencies. Internet Appendix Table A.2 shows that more than one quarter of the datasets on Datarade are updated at an intraday frequency (real time, secondly, or minutely), while more than 40% are updated at least daily. Such frequent information updates should

induce more frequent portfolio adjustments.

To examine this hypothesis, we consider three complementary measures of trading behavior. First, we measure portfolio turnover as the absolute change in the dollar value of a position after adjusting for price changes:

$$Turnover_{ijt} = 100 \times \frac{|MV_{ijt} - MV_{ij,t-1} - N_{ij,t-1}\,(P_{jt} - P_{j,t-1})|}{\frac{1}{2}\,(MV_{ijt} + MV_{ij,t-1})}, \tag{6}$$

where MV_{ijt} denotes the dollar value of investor i's position in stock j at the end of quarter t, N_{ijt} is the corresponding number of shares held, and P_{jt} is the quarter-end stock price. The numerator measures the absolute dollar amount traded during the quarter after removing the passive change in portfolio value attributable to stock price movements, while the denominator scales trading activity by the average dollar value of the position over the quarters.

Second, we examine churn, defined as an indicator equal to one when an investor either initiates or completely liquidates a position during the quarter. Third, we study holding duration, measured as the number of consecutive quarters that an investor continuously holds a position. Recall that our sample includes all stocks belonging to an investor's investment universe, defined following Koijen and Yogo (2019) as the set of stocks currently held or held at any point during the previous eleven quarters. Consequently, stocks that belong to the investment universe but are not currently held remain in the sample with a portfolio weight of zero. Retaining these zero holdings allows us to identify complete exits when constructing the churn and holding-duration measures.

Table V reports the results. We find that following increases in alternative data coverage, hedge funds trade covered stocks significantly more actively than other institutional investors. Portfolio turnover increases by approximately 2 percentage points, while the probability of completely entering or exiting a position rises by 0.6 percentage points. These results indicate that hedge funds rebalance their positions more aggressively once firms become covered by commercially available alternative datasets.

We also find that hedge fund positions in covered firms become significantly shorter-lived. On average, holding durations decline by approximately two quarters following coverage. Taken together, these findings suggest that alternative data affect not only the composition

of hedge fund portfolios but also the frequency with which these portfolios are updated, consistent with alternative data generating a richer flow of firm-specific information that induces more frequent trading.

D Are the Effects Stronger for Quantitative Hedge Funds?

The previous analysis treats hedge funds as a homogeneous group. In practice, however, hedge funds differ in their investment processes and their ability to exploit large and complex datasets. If commercial alternative data primarily generate value through quantitative analysis, such as machine learning, one would expect the portfolio response documented above to be strongest among quantitative hedge funds.

Directly observing whether a hedge fund follows a quantitative investment strategy is not possible. We therefore construct a proxy using investment advisers' regulatory disclosures in Form ADV Part 2A. Registered investment advisers are required to describe their methods of analysis and investment strategies in Item 8 ("Methods of Analysis, Investment Strategies and Risk of Loss"). Following Chen et al. (2026), we collect all Form ADV Part 2A filings from the SEC EDGAR database beginning in 2012, when significant revisions to Form ADV implementing the Dodd–Frank Act became effective. For each adviser, we extract Item 8 and retain the last filing submitted in each calendar year. We then use FactSet's entity link table to match hedge funds in our ownership database to the corresponding SEC CRD identifier.

We classify a hedge fund as quantitatively oriented using a keyword-based approach. Specifically, we count the occurrence of terms associated with quantitative investing ("algorithm", "approx", "computer", "model", "statistical", "system", "data", "automated", "quantitative", and "math") and subtract the occurrence of terms associated with discretionary investing ("discretionary", "fundamental", and "qualitative"). A hedge fund is classified as quantitative in a given year whenever the resulting score is strictly positive.

To examine whether quantitative hedge funds respond differently to alternative data coverage, we augment Equation (4) by interacting the treatment indicator with an indicator for quantitative hedge funds. The coefficient on this additional interaction captures the incremental response of quantitative hedge funds relative to other hedge funds. Since quantitative hedge funds are a subset of hedge funds, the total effect for quantitative hedge

funds is given by the sum of the coefficients on the hedge fund interaction and the quantitative hedge fund interaction.

Internet Appendix Table D.3 reports the results. Consistent with the hypothesis that commercial alternative data are exploited to larger extent by investors with stronger data-processing capabilities, quantitative hedge funds exhibit significantly greater trading activity following increases in alternative data coverage. Relative to other hedge funds, they increase portfolio turnover by an additional 1.1 percentage points, are twice more likely to enter or exit covered stocks, and reduce their holding horizons by a further 0.7 quarters. In contrast, their increase in portfolio weights is statistically indistinguishable from that of other hedge funds. Overall, these results suggest that quantitative hedge funds trade trade more actively following newly available alternative data, relative to other hedge funds.

V Alternative Data and Firm-Level Ownership Composition

The previous section shows that alternative data coverage induces heterogeneous portfolio reallocations across institutional investors. We now ask whether these portfolio adjustments translate into persistent changes in firms' shareholder bases. In particular, we examine whether the increase in hedge fund trading documented at the investor–stock level ultimately changes the ownership composition of covered firms. Such a translation is not mechanical. The holding-level results establish that hedge funds increase their demand for covered stocks relative to other stocks, compared to other investors, but they do not imply that hedge funds ultimately own a larger fraction of these firms on average.

A Empirical Strategy

Studying changes in firm-level ownership is challenging because the timing of alternative data coverage may coincide with other firm-level developments that also affect institutional ownership. Consequently, a simple comparison of ownership before and after firms become highly covered by alternative data may confound the effect of coverage with contemporaneous changes in firms' characteristics. To alleviate these concerns, we combine the stacked difference-

in-differences estimator of Cengiz et al. (2019) and Gormley and Matsa (2011) with a matching procedure. This approach accommodates staggered treatment timing while ensuring that treated firms are compared with firms exhibiting similar pre-treatment characteristics.

Specifically, for each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage. We then construct a treatment cohort consisting of the treated firm and its five nearest control firms. Control firms are selected using nearest-neighbor matching based on standardized pre-treatment market capitalization and the ownership shares held by each investor category (investment advisors, hedge funds, long-term investors, private banks, and brokers).[20] Matching on ownership composition ensures that treated and control firms have similar shareholder bases prior to treatment. The matching variables are measured two years before treatment.

For each treatment cohort, we retain observations from two years before to three years after treatment and stack all cohorts into a single estimation sample. We then estimate

$$Y_{j,t,c} = \beta \left\{ Covered \times Post \right\}_{j,t} + \Gamma' X_{j,t} + \alpha_{j \times c} + \gamma_{t \times c} + \varepsilon_{j,t,c}, \tag{7}$$

where j indexes firms, t quarters, and c treatment cohorts. The dependent variable $Y_{j,t,c}$ measures firm-level ownership outcomes, such as the ownership share of each investor type. The indicator $Covered \times Post$ equals one after firm j enters the highest tercile of alternative data coverage, zero otherwise. The vector $X_{j,t}$ includes log market capitalization, market beta, leverage (debt-to-assets ratio), dividend-to-book equity, log book equity, and lagged one-quarter stock returns. Standard errors are two-way clustered by firm and year-quarter.

The specification includes firm×cohort fixed effects, $\alpha_{j \times c}$, which absorb all time-invariant differences between treated firms and their matched controls within each cohort, and year-quarter×cohort fixed effects, $\gamma_{t \times c}$, which absorb aggregate shocks common to all firms in a given treatment cohort. Identification therefore comes from comparing changes in ownership composition between treated firms and their matched controls following alternative data coverage.

[20] Specifically, the distance between a potential control and a treated firm is computed as the sum of the absolute differences in matching variables.

B Changes in Ownership Composition

Table VI reports the estimation results of regression (7) where the dependent variable is the fraction of a firm's ownership held by each investor type.

Column (1) shows that overall institutional ownership remains largely unchanged following increases in alternative data coverage. Thus, alternative data does not appear to increase institutional ownership per se. However, Columns (2) to (6) show that it changes the composition of institutional ownership across investor types.

The most striking result appears in Column (4). Following entry into the highest tercile of alternative data coverage, the ownership share held by hedge funds increases by 0.37 percentage points (p.p.), a statistically and economically significant effect representing 13% relative to the sample median reported in Table I Panel B (2.8 p.p.). In contrast, the ownership shares of brokers, long-term investors, and private banks all decline, although these effects are statistically insignificant. The negative coefficients for these investor groups, together with the absence of any change in aggregate institutional ownership, indicate that the increase in hedge fund ownership primarily reflects a reallocation of ownership away from other institutional investors rather than an expansion of institutional ownership overall.

To better assess the economic magnitude of the ownership reallocation, Internet Appendix Table D.4 examines whether alternative data coverage increases the likelihood that hedge funds hold economically meaningful ownership stakes. Specifically, we estimate the stacked difference-in-differences specification in equation (7) using indicator variables equal to one if hedge fund ownership exceeds a series of thresholds ranging from 0% to 10% of the firm's equity. We find that alternative data significantly raises the probability that hedge funds hold ownership between 2% and 5% of a firm's equity, with positive and marginally significant effects even for ownership stakes exceeding 10%. These results suggest that alternative data coverage is associated with economically meaningful reallocations of ownership toward hedge funds.

C Displacement of Long-Term Investors

The previous subsection shows that alternative data coverage is associated with an increase in hedge fund ownership. We now examine whether this reallocation occurs at the expense of long-term shareholders. While the investor-type classification identifies pension funds and insurance companies as traditional long-term investors, investment horizon may also vary substantially within each investor category. We therefore adopt a direct approach that classifies investors according to the duration for which they have continuously held a firm's stock.

Table VII reports the results. The dependent variable is the fraction of institutional ownership held by investors whose positions have been maintained continuously for at least two, four, six, eight, ten, and twelve consecutive quarters. The results reveal a clear pattern. The estimated effect of alternative data coverage becomes increasingly negative as the holding-duration threshold increases. Treated firms experiencing greater alternative data coverage exhibit a statistically and economically significant decline in ownership by investors that have held the stock for at least eight quarters, relative to control firms. The estimated effect reaches approximately one percentage point for investors with holding horizons of ten to twelve quarters.

Taken together, these results indicate that the increase in hedge fund ownership documented in the previous subsection is accompanied by a displacement of long-term shareholders. Rather than attracting additional institutional capital, alternative data coverage appears to alter the composition of firms' shareholder bases by reallocating ownership from patient investors toward institutions that trade more actively and exploit newly available information. This shift in ownership composition motivates our subsequent analysis of the implications of alternative data for stock price dynamics and market fragility.

VI Market Consequences of Alternative Data

The previous section shows that the expansion of commercial alternative data reshapes firms' ownership structure by reallocating ownership away from long-term investors and toward hedge funds. We now examine whether this change in ownership has implications for financial

markets and stock prices.

On the one hand, alternative data provide investors with new sources of firm-specific information that can be incorporated more rapidly into prices. If so, stocks becoming more covered by alternative data should exhibit greater information efficiency (as shown in Zhu, 2019; Bonelli and Foucault, 2025, using alternative data tracking retailers' parking lots), with returns reflecting more firm-specific information and less comovement with aggregate market or industry-wide shocks.

On the other hand, the investors best equipped to exploit alternative data are also among the most active market participants. As documented in the previous section, greater alternative data coverage attracts hedge funds while displacing long-term investors, resulting in a shareholder base characterized by shorter investment horizons and higher trading intensity. If these investors withdraw capital more aggressively during periods of market downturns, the resulting ownership structure may increase firms' exposure to fluctuations in capital supply, amplifying downside risk and stock price fragility during these periods.

We examine these two channels. We first study whether alternative data coverage reduces stock return comovement with the market, industry, and firms similarly covered by alternative data, consistent with more incorporation of firm-specific information into prices. We then examine whether the shift toward hedge fund ownership increases firms' downside return sensitivity and induces larger hedge fund outflows during periods of market stress.

A Return Comovements

Our ownership results suggest that alternative data shift ownership toward hedge funds. If these investors process information contained in alternative data, stock returns should become less sensitive to aggregate market and industry movements and instead reflect a greater amount of firm-specific information. To examine this prediction, we study whether alternative data coverage affects stock return comovements. Following Malceniece et al. (2019), we measure return comovement using the R^2 from daily regressions estimated separately for each stock and quarter using data from the CRSP database. Specifically, we estimate:

$$r_{i,t} = \alpha_i + \sum_{k=-1}^{1} \beta_k r_{g,t+k} + \varepsilon_{i,t}, \tag{8}$$

where $r_{i,t}$ denotes the daily return of stock i, and $r_{g,t}$ denotes the equally weighted return of a reference group excluding stock i.[21]

We consider three reference groups. First, we compute market comovement using all publicly traded firms. Second, we compute industry comovement using firms belonging to the same industry.[22] Finally, we compute a measure of alternative-data comovement by constructing reference portfolios consisting of firms with similar alternative data exposure. Specifically, firms are grouped into terciles according to their alternative data coverage measure, as described in Section III.C, and the reference portfolio is formed as the equally weighted return of firms within the same tercile.

The R^2 coefficient from equation (8) measures the fraction of a stock's return variation explained by common movements in the corresponding reference portfolio. Higher values therefore indicate greater return comovement and less firm-specific return variation. Following Morck et al. (2000); Hameed et al. (2010); Malceniece et al. (2019), we apply the logit transformation

$$Logit(R^{2,g}_{i,t}) = \ln\left(\frac{R^2}{1-R^2}\right), \qquad g \in \{Market, Industry, Data\}, \tag{9}$$

to transform the R^2 statistic into an unbounded variable. We then estimate the stacked difference-in-differences specification introduced in equation (7), replacing the dependent variable with each of the three comovement measures.

Table VIII reports the results. Across all specifications, treated firms experiencing higher alternative data coverage exhibit significantly lower return comovement with the market, their industry peers, and firms with similar alternative data exposure, relative to control firms. The estimated effects are economically meaningful. Relative to control firms, market comovement

[21] We include contemporaneous, lagged, and lead returns following Malceniece et al. (2019) to accounts for nonsynchronous trading.

[22] We use the stock's Fama-French 48-industry classification to be consistent with our alternative data coverage measure.

declines by 1.3 percentage points following coverage, while industry and alternative-data comovement decline by 1.2 and 0.8 percentage points, respectively. The results are similar when using the logit transformation of the corresponding R^2 measures. These findings indicate that, post coverage, stock returns become less driven by common market and industry factors following increases in alternative data coverage.

One interpretation is that the growing presence of data-savvy investors, particularly hedge funds, facilitates the incorporation of firm-specific information into prices. Therefore, if the reduction in return comovement is driven by the trading activity of these informed investors, then it should be attenuated during periods when hedge funds withdraw capital. The withdrawal of informed capital should temporarily reduce the incorporation of firm-specific information into prices, leading covered firms' returns to move more closely with the market (and their industry peers). We test this prediction in the next subsection.

B Fragility During Market Downturns

B.1 Comovements During Market Downturns

Internet Appendix Table D.5 reports the relation between hedge fund ownership and our three measures of return comovement. Across all specifications, greater hedge fund ownership is associated with significantly lower return comovement with the market, industry peers, and firms with similar levels of alternative data coverage. These results are consistent with the view that hedge funds incorporate firm-specific information into prices more aggressively than other institutional investors, thereby increasing the firm-specific component of stock returns.

If hedge funds are indeed responsible for these informational efficiency gains, one would expect the effect to weaken precisely when these investors withdraw capital. During periods of market stress, hedge funds may face funding constraints, deleveraging pressures, or heightened risk aversion, leading them to unwind positions and reduce the incorporation of firm-specific information into prices. As a result, covered firms' returns should become more closely synchronized with aggregate market movements during market downturns, despite exhibiting lower comovement on average.

To examine this prediction, we extend our baseline specification by interacting the coverage effect with measures of aggregate market conditions. Specifically, we estimate separate specifications interacting $Covered \times Post$ with (i) the contemporaneous market return in the quarter, (ii) an indicator for quarters in which the market return falls in the bottom decile of its historical distribution ("Low Market Return"), and (iii) an indicator for quarters in which the market return belongs to the top decile ("High Market Return").

Table IX reports the results. Consistent with the previous subsection, the average effect of alternative data coverage remains negative across all specifications. However, this reduction is significantly attenuated during periods of market downturns. The interaction between $Covered \times Post$ and the low-market-return indicator is positive and statistically significant for all three measures of return comovement. In contrast, we find no evidence that the effect differs during periods of high market returns.

These results suggest that during normal market conditions, covered firms exhibit substantially greater firm-specific return variation after coverage. During market downturns, however, these gains largely disappear, with covered firms' returns moving more closely with the market, their industry peers, and firms exposed to similar alternative data.

One interpretation is that the informational efficiency generated by alternative data relies on the participation of sophisticated investors, such as hedge funds. When these investors reduce risk or face funding constraints during market downturns, firm-specific information is incorporated less aggressively into prices, causing covered firms' returns to become more synchronized with broader market movements. To confirm this interpretation, we examine below whether hedge funds reduce their holdings of covered firms more aggressively during periods of market stress and the implications for firms' stock returns.

B.2 Hedge Fund Trading and Implications for Stock Returns

The previous subsection shows that the reduction in return comovement associated with alternative data coverage largely disappears during market downturns. A natural explanation is that the informational advantages generated by alternative data depend on the presence of sophisticated investors actively processing these new information sources. If these investors reduce their positions during market downturns, the comovement effect documented above

should weaken precisely when market conditions deteriorate.

To investigate this mechanism directly, we examine whether hedge funds adjust their trading behavior differentially during periods of market stress. We estimate the same specification as in the previous subsection, replacing return comovement with hedge fund net trading, measured as net purchases minus net sales scaled by shares outstanding. We simultaneously examine whether these trading patterns are associated with differences in realized stock returns.

Table X reports the results. Columns (1)–(3) show that alternative data coverage has little effect on stock returns during normal market conditions. However, when aggregate market returns fall into the bottom decile of their historical distribution, covered firms experience significantly lower returns relative to control firms. The interaction between $Covered \times Post$ and the low-market-return indicator is negative and statistically significant, whereas no comparable effect is observed during periods of strong market performance.

Columns (4)–(6) provide direct evidence on the underlying trading mechanism. During normal market conditions, hedge fund net trading in covered firms is not significantly different post coverage, from that in control firms. In contrast, during market downturns and post treatment, hedge funds become significant more net sellers of covered firms. The interaction between $Covered \times Post$ and the low-market-return indicator is negative and economically meaningful, indicating that, post coverage, hedge funds withdraw capital disproportionately from covered firms when aggregate market conditions deteriorate, relative to control firms. Again, we find no corresponding effect during periods of high market returns.

Taken together, these results provide a coherent interpretation of the earlier evidence on return comovement: Alternative data attract sophisticated investors who actively trade on these signals, thereby improving the incorporation of firm-specific information into prices, which reduces return comovement during normal times. However, this change in the composition of firms' shareholder base has additional implications: Because hedge funds maintain shorter investment horizons and adjust their portfolios more aggressively during periods of market stress, they disproportionately reduce their positions in covered firms when market conditions deteriorate. Consequently, the informational benefits associated with alternative data weaken precisely during these periods, and covered firms become simultaneously more exposed to aggregate shocks and more vulnerable to downside market

returns.

Overall, the evidence suggests that commercial alternative data have two opposing effects on financial markets. On the one hand, they facilitate the incorporation of firm-specific information into stock prices by sophisticated investors. On the other hand, they reallocate ownership toward a more active and less stable investor base. This ownership channel creates a source of financial fragility, as the withdrawal of informed hedge fund capital during market downturns amplifies the sensitivity of covered firms to aggregate market downturns.

VII Conclusion

Recent technological innovations in information production are transforming firms' informational environment by enabling investors to obtain information from increasingly abundant data. We study one important manifestation of this transformation: the rapid expansion of commercial alternative data generated outside firms. We develop a novel text-based measure of firms' exposure to these new information sources. We show that greater alternative data coverage reshapes firms' ownership structures by increasing hedge funds' ownership, raising their trading activity, shortening their holding horizons, and displacing long-term shareholders. These findings identify ownership reallocation as a key channel through which innovations in information production affect firms.

This ownership reallocation has both beneficial and adverse consequences for financial markets. On the one hand, stocks of highly covered firms exhibit lower return comovement with the market and their industry, consistent with greater incorporation of firm-specific information into prices. On the other hand, these informational gains come at the cost of greater fragility. During periods of market stress, hedge funds reduce their holdings of highly covered firms more aggressively, making these firms' stock prices more sensitive to market downturns. Thus, innovations in information production can simultaneously improve informational efficiency and increase financial fragility by changing the composition of firms' shareholder bases.

More broadly, our findings suggest that the consequences of technological progress in information production extend beyond investors' information sets. By reshaping firms' ownership structures, these innovations also alter the way information is incorporated into

prices and the resilience of financial markets to aggregate shocks. There might be other effects. For instance, the displacement of long-term shareholders might lead to increased short-termism with consequences for corporate investment and governance. We leave the analysis of these effects for future research.

References

Aghion, P., Van Reenen, J., and Zingales, L. (2013). Innovation and institutional ownership. The American Economic Review, 103(1):277–304.

Becht, M., Franks, J., Grant, J., and Wagner, H. F. (2017). Returns to hedge fund activism: An international study. The Review of Financial Studies, 30(9):2933–2971.

Becht, M., Franks, J., Mayer, C., and Rossi, S. (2009). Returns to shareholder activism: Evidence from a clinical study of the hermes uk focus fund. The Review of Financial Studies, 22(8):3093–3129.

Begenau, J., Farboodi, M., and Veldkamp, L. (2018). Big data in finance and the growth of large firms. Journal of Monetary Economics, 97:71–87.

Ben-David, I., Franzoni, F., Moussawi, R., and Sedunov, J. (2021). The granular nature of large institutional investors. Management Science, 67(11):6629–6659.

Bhojraj, S., Cho, Y. J., and Yehuda, N. (2012). Mutual fund family size and mutual fund performance: The role of regulatory changes. Journal of Accounting Research, 50(3):647–684.

Bonelli, M. (2026). Data-driven investors. The Review of Financial Studies, 39(7):1909–1969.

Bonelli, M. and Foucault, T. (2025). Does big data devalue traditional expertise? evidence from active funds. SSRN Electronic Journal.

Boyson, N. M., Stahel, C. W., and Stulz, R. M. (2010). Hedge fund contagion and liquidity shocks. The Journal of Finance, 65:1789–1816.

Brav, A., Jiang, W., Partnoy, F., and Thomas, R. (2008). Hedge fund activism, corporate governance, and firm performance. The Journal of Finance, 63(4):1729–1775.

Brunnermeier, M. K. and Pedersen, L. H. (2009). Market liquidity and funding liquidity. The Review of Financial Studies, 22(6):2201–2238.

Bushee, B. J. (1998). The influence of institutional investors on myopic R&D investment behavior. The Accounting Review, 73(3):305–333.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). From man vs. machine to man + machine: The art and ai of stock analyses. Journal of Financial Economics, 160:103910.

Cella, C., Ellul, A., and Giannetti, M. (2013). Investors' horizons and the amplification of market shocks. The Review of Financial Studies, 26(7):1607–1648.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. The Quarterly Journal of Economics, 134(3):1405–1454.

Chen, S., Sialm, C., and Xu, D. X. (2026). The growth and performance of artificial intelligence in asset management. Technical report, National Bureau of Economic Research.

Chen, T., Dong, H., and Lin, C. (2020). Institutional shareholders and corporate social responsibility. Journal of Financial Economics, 135(2):483–504.

Chi, F., Hwang, B.-H., and Zheng, Y. (2024). The use and usefulness of big data in finance: Evidence from financial analysts. Management Science.

Cong, L. W., Xie, D., and Zhang, L. (2021). Knowledge accumulation, privacy, and growth in a data economy. Management Science, 67(10):6480–6492.

Cremers, M., Pareek, A., and Sautner, Z. (2020). Short-term investors, long-term investments, and firm value: Evidence from russell 2000 index inclusions. Management Science, 66(10):4535–4551.

Derrien, F., Kecskés, A., and Thesmar, D. (2013). Investor horizons and corporate policies. Journal of Financial and Quantitative Analysis, 48(6):1755–1780.

Dessaint, O., Foucault, T., and Frésard, L. (2024). Does alternative data improve financial forecasting? the horizon effect. The Journal of Finance, 79(3):2237–2287.

Dow, J., Han, J., and Sangiorgi, F. (2024). The short-termism trap: Catering to informed investors with limited horizons. Journal of Financial Economics, 159:103884.

Dugast, J. and Foucault, T. (2025). Equilibrium data mining and data abundance. The Journal of Finance, pages 211–258.

Eisfeldt, A. L., Schubert, G., Taska, B., and Zhang, M. B. (2026). Generative ai and firm values. Journal of Finance. Forthcoming.

Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2022). Where has all the data gone? The Review of Financial Studies, 35(7):3101–3138.

Farboodi, M. and Veldkamp, L. (2026). A model of the data economy. The Review of Economic Studies, page rdag074.

Gao, M. and Huang, J. (2019). Informing the market: The effect of modern information technologies on information production. The Review of Financial Studies, 33:1367–1411.

Glossner, S., Matos, P., Ramelli, S., and Wagner, A. F. (2025). Do institutional investors stabilize equity markets in crisis periods? evidence from covid-19. Management Science. Forthcoming.

Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big Data in Finance. The Review of Financial Studies, 34(7):3213–3225.

Gormley, T. A. and Matsa, D. A. (2011). Growing out of trouble? corporate responses to liability risk. The Review of Financial Studies, 24(8):2781–2821.

Green, T. C., Huang, R., Wen, Q., and Zhou, D. (2019). Crowdsourced employer reviews and stock returns. Journal of Financial Economics, 134:236–251.

Greenwood, R. and Thesmar, D. (2011). Stock price fragility. Journal of Financial Economics, 102(3):471–490.

Gârleanu, N. and Pedersen, L. H. (2007). Liquidity and risk management. American Economic Review, 97(2):193–197.

Hameed, A., Kang, W., and Viswanathan, S. (2010). Stock market declines and liquidity. The Journal of finance, 65(1):257–293.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of political economy, 124(5):1423–1465.

Hoberg, G. and Phillips, G. M. (2025). Scope, scale, and concentration: The 21st-century firm. The Journal of Finance, 80(1):415–466.

Huang, J. (2018). The customer knows best: The investment value of consumer opinions. Journal of Financial Economics, 128:164–182.

Katona, Z., Painter, M. O., Patatoukas, P. N., and Zeng, J. (2024). On the capital market consequences of big data: Evidence from outer space. Journal of Financial and Quantitative Analysis, 59(4):1473–1501.

Katona, Z., Painter, M. O., Patatoukas, P. N., and Zeng, J. (2025). On the capital market consequences of big data: Evidence from outer space. Journal of Financial and Quantitative Analysis, 60(2):551–579.

Klein, A. and Zur, E. (2009). Entrepreneurial shareholder activism: Hedge funds and other private investors. Journal of Finance, 64(1):187–229.

Koijen, R. S., Richmond, R. J., and Yogo, M. (2024). Which investors matter for equity valuations and expected returns? Review of Economic Studies, 91(4):2387–2424.

Koijen, R. S. and Yogo, M. (2019). A demand system approach to asset pricing. Journal of Political Economy, 127(4):1475–1515.

Kruttli, M. S., Monin, P. J., Petrasek, L., and Watugala, S. W. (2025). Ltcm redux? hedge fund treasury trading, funding fragility, and risk constraints. Journal of Financial Economics, 169:104017.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.

Liu, C. Y., Low, A., Masulis, R. W., and Zhang, L. (2020). Monitoring the monitor: Distracted institutional investors and board governance. The Review of Financial Studies, 33(10):4489–4531.

Liu, E., Ma, S., and Veldkamp, L. (2025). Data sales and data dilution. Journal of Financial Economics, 169:104053.

Malceniece, L., Malcenieks, K., and Putniņš, T. J. (2019). High frequency trading and comovement in financial markets. Journal of Financial Economics, 134(2):381–399.

Mihet, R., Rishabh, K., and Gomes, O. (2026). Is it ai or data that drives firm market power? Journal of Monetary Economics, 157:103878.

Morck, R., Yeung, B., and Yu, W. (2000). The information content of stock markets: why do emerging markets have synchronous stock price movements? Journal of financial economics, 58(1-2):215–260.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. The Journal of Finance, 52(1):35–55.

Sun, Q., Zhang, H., Liu, J., Mao, X., Liu, J., Ren, K., and Hu, H. (2026). Dadada: A dataset for data pricing in data marketplaces. arXiv preprint arXiv:2607.08785.

Zhu, C. (2019). Big data as a governance mechanism. Review of Financial Studies, 32(5):2021–2061.

Tables

Table I: Sample Descriptive Statistics

This table reports summary statistics for the main variables used throughout the paper. Panel A reports statistics at the investor–stock–quarter level. The sample covers 11,077 institutional investors and 10,955 stocks between 2000-Q1 and 2023-Q4. Panel B reports summary statistics at the stock–quarter level. Ownership shares are reported in percentage points.

Panel A: Holding (Investor-Stock-Quarter) Level

	Obs	Mean	Sd	5%	25%	50%	75%	95%
Weight (%)	2.77e+07	0.350	1.522	0.000	0.004	0.030	0.195	1.707
Excess Weight (%)	2.77e+07	0.021	1.616	-1.051	-0.099	-0.010	0.021	1.105
Dollar Turnover (%)	2.77e+07	52.171	73.823	0.000	1.249	11.157	73.141	200.000
Churn	2.77e+07	0.153	0.360	0.000	0.000	0.000	0.000	1.000
Holding Duration (Quarters)	2.77e+07	12.833	15.732	1.000	3.000	7.000	17.000	47.000
Covered × Post	2.77e+07	0.212	0.409	0.000	0.000	0.000	0.000	1.000
Investment Advisor	2.77e+07	0.617	0.486	0.000	0.000	1.000	1.000	1.000
Broker	2.77e+07	0.075	0.264	0.000	0.000	0.000	0.000	1.000
Hedge Fund	2.77e+07	0.102	0.303	0.000	0.000	0.000	0.000	1.000
Long-term	2.77e+07	0.061	0.239	0.000	0.000	0.000	0.000	1.000
Private Bank	2.77e+07	0.145	0.352	0.000	0.000	0.000	0.000	1.000

Panel B: Stock-Quarter Level

	Obs	Mean	Sd	5%	25%	50%	75%	95%
Mkt Cap (mio)	125,672	5,505.000	40375.119	16.063	104.685	470.460	2,154.785	18219.781
Beta	125,672	0.984	0.924	-0.199	0.444	0.886	1.404	2.538
Debt/Asset	125,672	0.217	0.202	0.000	0.034	0.172	0.352	0.600
Dividends/Book Equity	125,672	0.031	1.054	0.000	0.000	0.000	0.010	0.035
Book Equity (mio)	125,672	1,998.718	9,109.939	10.743	63.520	243.817	933.566	7,365.414
Stock Return	125,672	0.032	0.456	-0.374	-0.109	0.011	0.133	0.458
Covered × Post	125,672	0.167	0.373	0.000	0.000	0.000	0.000	1.000
Advisor Ownership	125,672	27.957	19.643	0.343	9.575	28.065	43.730	59.774
Broker Ownership	125,672	1.038	1.582	0.000	0.184	0.654	1.313	3.348
Hedge Fund Ownership	125,672	5.621	7.654	0.000	0.794	2.844	7.325	20.905
Long-Term Ownership	125,672	1.678	2.078	0.000	0.236	1.093	2.549	4.805
Private Bank Ownership	125,672	2.212	3.853	0.000	0.195	1.095	2.623	8.275
Institutional Ownership	125,672	55.191	31.462	4.050	26.551	59.558	83.414	100.000

Table II: Determinants of Alternative Data Coverage

This table examines the relation between firms' alternative data coverage and observable firm characteristics. The dependent variable in all specifications is the alternative data coverage measure (standardized over the full sample to have zero mean and unit variance). The independent variables are the logarithm of market capitalization, total assets, and book-to-market ratio. Columns (1)–(4) report regressions including year-quarter fixed effects. Columns (5)–(8) additionally include firm fixed effects, so the estimated coefficients are identified from within-firm changes over time. Standard errors are two-way clustered by firm and year-quarter.

	Coverage							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log(Mkt Cap)	0.017**			-0.209***	0.013			0.025
	(0.007)			(0.021)	(0.009)			(0.016)
Log(Assets)		0.072***		0.244***		0.027**		0.002
		(0.006)		(0.020)		(0.010)		(0.017)
Log(Book-to-Market)			0.098***	-0.104***			0.021**	0.034**
			(0.016)	(0.018)			(0.009)	(0.015)
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	125,515	125,515	125,515	125,515	125,515	125,515	125,515	125,515
R^2	0.11	0.13	0.12	0.16	0.63	0.63	0.63	0.63

Table III: Alternative Data Coverage and Institutional Portfolio Holdings

This table reports the effect of alternative data coverage on institutional investors' portfolio holdings. Regressions are estimated at the investor–stock–quarter level. The dependent variable is the stock's portfolio weight in Columns (1) and (2), and the stock's excess portfolio weight relative to its market weight in Columns (3) and (4). The variable *Covered* × *Post* is an dummy equal to one after a firm enters the highest tercile of alternative data coverage. All specifications include stock fixed effects and investor × year-quarter fixed effects. Standard errors are three-way clustered by investor, stock, and year-quarter.

	Weight (%)		Excess Weight (%)	
	(1)	(2)	(3)	(4)
Covered × Post	-0.018**	-0.016*	0.012	0.019
	(0.009)	(0.009)	(0.016)	(0.018)
Log(Mkt Cap)		0.141***		-0.146*
		(0.029)		(0.077)
Beta		-0.001		-0.021**
		(0.005)		(0.009)
Debt/Asset		0.051		-0.364**
		(0.053)		(0.162)
Dividends/Book Equity		-0.004***		0.004
		(0.001)		(0.003)
Log(Book Equity)		-0.037***		0.004
		(0.011)		(0.024)
Stock Return (t-1)		0.008		0.047*
		(0.008)		(0.024)
Company FE	Yes	Yes	Yes	Yes
Investor × Year-Quarter FE	Yes	Yes	Yes	Yes
Observations	2.77e+07	1.95e+07	2.77e+07	1.95e+07
R^2	0.47	0.49	0.52	0.55

Table IV: Alternative Data Coverage and Portfolio Reallocation Across Investor Types

This table reports the effect of alternative data coverage on institutional investors' portfolio holdings. Regressions are estimated at the investor–stock–quarter level. The dependent variable is the stock's portfolio weight in Columns (1) and (2), and the stock's excess portfolio weight relative to its market weight in Columns (3) and (4). The variable $Covered \times Post$ is an indicator equal to one after a firm enters the highest tercile of alternative data coverage. The reported coefficient is the interaction between $Covered \times Post$ and the investor type. The omitted type in Columns (2) and (4) is Investment Advisor. All specifications include company × year-quarter fixed effects and institution × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

	Weight (%)		Excess Weight (%)	
	(1)	(2)	(3)	(4)
Covered × Post × Hedge Fund	0.033***	0.030***	0.033***	0.030***
	(0.011)	(0.010)	(0.011)	(0.010)
Covered × Post × Broker		-0.009***		-0.009***
		(0.003)		(0.003)
Covered × Post × Long-Term		0.004		0.004
		(0.006)		(0.006)
Covered × Post × Private Banking		-0.019***		-0.019***
		(0.006)		(0.006)
Company × Year-Quarter FE	Yes	Yes	Yes	Yes
Investor × Year-Quarter FE	Yes	Yes	Yes	Yes
Observations	2.77e+07	2.77e+07	2.77e+07	2.77e+07
R^2	0.48	0.48	0.56	0.56

Table V: Alternative Data Coverage and Trading Behavior Across Investor Types

This table reports the effect of alternative data coverage on institutional investors' trading behavior. Regressions are estimated at the investor–stock–quarter level. The dependent variable is the dollar turnover (in %) in Columns (1) and (2), the churn indicator (entry or exit) in Columns (3) and (4), and the holding duration in quarters in Columns (5) and (6). The variable *Covered* × *Post* is an indicator equal to one after a firm enters the highest tercile of alternative data coverage. The reported coefficient is the interaction between *Covered* × *Post* and the investor type. The omitted type in Columns (2), (4), and (6) is Investment Advisor. All specifications include company × year-quarter fixed effects and institution × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

	Dollar Turnover (%)		Churn		Holding Duration	
	(1)	(2)	(3)	(4)	(5)	(6)
Covered × Post × Hedge Fund	2.283***	2.160***	0.006***	0.006***	-1.994***	-1.932***
	(0.401)	(0.396)	(0.002)	(0.002)	(0.300)	(0.302)
Covered × Post × Broker		-0.592		-0.004**		0.479
		(0.379)		(0.002)		(0.341)
Covered × Post × Long-Term		-0.232		-0.000		1.631***
		(0.194)		(0.001)		(0.361)
Covered × Post × Private Banking		-0.337*		-0.001		-0.634***
		(0.178)		(0.001)		(0.197)
Company × Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Investor × Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2.77e+07	2.77e+07	2.77e+07	2.77e+07	2.77e+07	2.77e+07
R^2	0.41	0.41	0.38	0.38	0.55	0.55

Table VI: Alternative Data Coverage and Institutional Ownership Composition

This table reports the effect of alternative data coverage on firms' institutional ownership. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variable is the ownership share held by all institutional investors in Column (1), investment advisors in Column (2), brokers in Column (3), hedge funds in Column (4), long-term investors in Column (5), and private banks in Column (6). The variable *Covered* × *Post* is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. All specifications include company × cohort fixed effects, year-quarter × cohort fixed effects, and control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns. Standard errors are two-way clustered by company and year-quarter.

	IO (All)	Advisor	Broker	Hedge Fund	Long-Term	Private Bank
	(1)	(2)	(3)	(4)	(5)	(6)
Covered × Post	0.196	0.101	-0.035	0.372***	-0.026	-0.082
	(0.232)	(0.198)	(0.030)	(0.113)	(0.023)	(0.054)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.97	0.96	0.80	0.90	0.92	0.90

Table VII: Alternative Data Coverage and Long-Term Institutional Ownership

This table reports the effect of alternative data coverage on long-term institutional ownership. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variable is the fraction of institutional ownership held by investors that have continuously held the firm's stock for at least two, four, six, eight, ten, and twelve consecutive quarters in Columns (1)–(6), respectively. The variable *Covered* × *Post* is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. All specifications include company × cohort fixed effects, year-quarter × cohort fixed effects, and control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns. Standard errors are two-way clustered by company and year-quarter.

	IO ≥ 2-Q	IO ≥ 4-Q	IO ≥ 6-Q	IO ≥ 8-Q	IO ≥ 10-Q	IO ≥ 12-Q
	(1)	(2)	(3)	(4)	(5)	(6)
Covered × Post	-0.357	-0.397	-0.570	-0.828**	-1.063**	-1.049***
	(0.284)	(0.375)	(0.397)	(0.403)	(0.416)	(0.392)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.83	0.86	0.89	0.90	0.91	0.92

Table VIII: Alternative Data Coverage and Stock Return Comovement

This table reports the effect of alternative data coverage on stock return comovement. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variables are quarterly return comovement measures. Specifically, return comovement is measured as the R^2 from a daily regression of a firm's stock returns on the contemporaneous, lagged, and lead equal-weighted returns of all other firms in the reference group. The reference group is the entire market in Columns (1) and (2), firms in the same industry in Columns (3) and (4), and firms in the same alternative-data similarity tercile in Columns (5) and (6). Even-numbered columns report the logit transformation of the corresponding R^2 measure. The variable $Covered \times Post$ is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. All specifications include company × cohort fixed effects, year-quarter × cohort fixed effects, and control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns. Standard errors are two-way clustered by company and year-quarter.

	$R^{2,mkt}$	$Logit(R^{2,mkt})$	$R^{2,ind}$	$Logit(R^{2,ind})$	$R^{2,data}$	$Logit(R^{2,data})$
	(1)	(2)	(3)	(4)	(5)	(6)
Covered × Post	-0.013***	-0.072***	-0.012***	-0.069***	-0.008***	-0.052**
	(0.003)	(0.022)	(0.003)	(0.022)	(0.003)	(0.022)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.77	0.75	0.80	0.77	0.77	0.75

Table IX: Alternative Data Coverage, Return Comovement, and Market Downturns

This table reports whether the effect of alternative data coverage on stock return comovement varies with aggregate market conditions. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variables are quarterly return comovement measures. Specifically, return comovement is measured as the R^2 from a daily regression of a firm's stock returns on the contemporaneous, lagged, and lead equal-weighted returns of all other firms in the reference group. The reference group is the entire market in Columns (1)–(3), firms in the same industry in Columns (4)–(6), and firms belonging to the same alternative-data similarity tercile in Columns (7)–(9). The variable *Covered* × *Post* is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. The reported triple-interaction coefficients measure whether the effect of alternative data coverage on return comovement varies with contemporaneous market returns in Columns (1), (4), and (7), periods of low market returns in Columns (2), (5), and (8), and periods of high market returns in Columns (3), (6), and (9). Low (High) Market Return is an indicator equal to one if the quarterly market return falls in the bottom (top) decile of the empirical distribution of quarterly market returns over the sample period. All specifications control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns, and include company × cohort fixed effects and year-quarter × cohort fixed effects. Standard errors are two-way clustered by company and year-quarter.

	$R^{2,mkt}$			$R^{2,ind}$			$R^{2,data}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Covered × Post	-0.012***	-0.014***	-0.012***	-0.011***	-0.014***	-0.011***	-0.007**	-0.010***	-0.008**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Covered × Post × Market Return	-0.055***			-0.065***			-0.046***		
	(0.016)			(0.019)			(0.017)		
Covered × Post × Low Market Return		0.013***			0.012**			0.012***	
		(0.004)			(0.005)			(0.005)	
Covered × Post × High Market Return			-0.002			-0.009			-0.003
			(0.006)			(0.006)			(0.006)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.77	0.77	0.77	0.80	0.80	0.80	0.77	0.77	0.77

Table X: Alternative Data Coverage, Stock Returns, and Hedge Fund Trading During Market Downturns

This table reports whether the effects of alternative data coverage on stock returns and hedge fund trading vary with aggregate market conditions. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variable is the firm's quarterly stock return in Columns (1)–(3) and hedge fund net trading, measured as net purchases minus net sales scaled by shares outstanding, in Columns (4)–(6). The variable *Covered* × *Post* is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. The reported triple-interaction coefficients measure whether the effect of alternative data coverage varies with contemporaneous market returns in Columns (1) and (4), periods of low market returns in Columns (2) and (5), and periods of high market returns in Columns (3) and (6). Low (High) Market Return is an indicator equal to one if the quarterly market return falls in the bottom (top) decile of the empirical distribution of quarterly market returns over the sample period. All specifications control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns, and include company × cohort fixed effects and year-quarter × cohort fixed effects. Standard errors are two-way clustered by company and year-quarter.

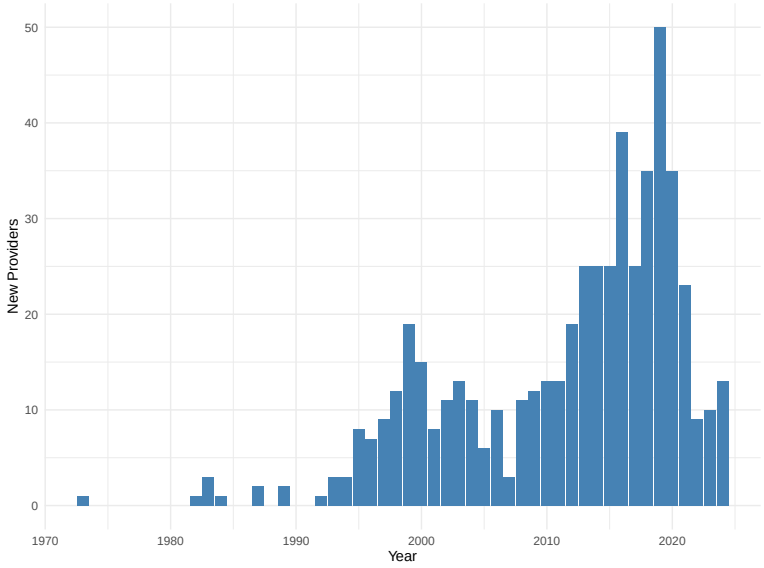
	Stock Return			Hedge Fund Net Trading (% Shares Outstanding)		
	(1)	(2)	(3)	(4)	(5)	(6)
Covered × Post	-0.003	0.004	-0.003	0.043	0.073	0.038
	(0.007)	(0.007)	(0.007)	(0.123)	(0.127)	(0.116)
Covered × Post × Market Return	0.106*			0.080		
	(0.055)			(0.564)		
Covered × Post × Low Market Return		-0.037***			-0.213**	
		(0.014)			(0.104)	
Covered × Post × High Market Return			0.024			0.065
			(0.015)			(0.223)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.55	0.55	0.55	0.38	0.38	0.38

Figures

Figure I: Growth in Commercial Alternative Data Providers

This figure plots the evolution of commercial alternative data providers listed on Datarade. Panel A reports the number of new providers entering the marketplace each year, while Panel B reports the cumulative number of providers over time. A provider is considered active once it is founded and offers at least one dataset whose coverage history has begun.

Panel A: Number of New Data Providers Each Year



Panel B: Cumulative Number of Data Providers

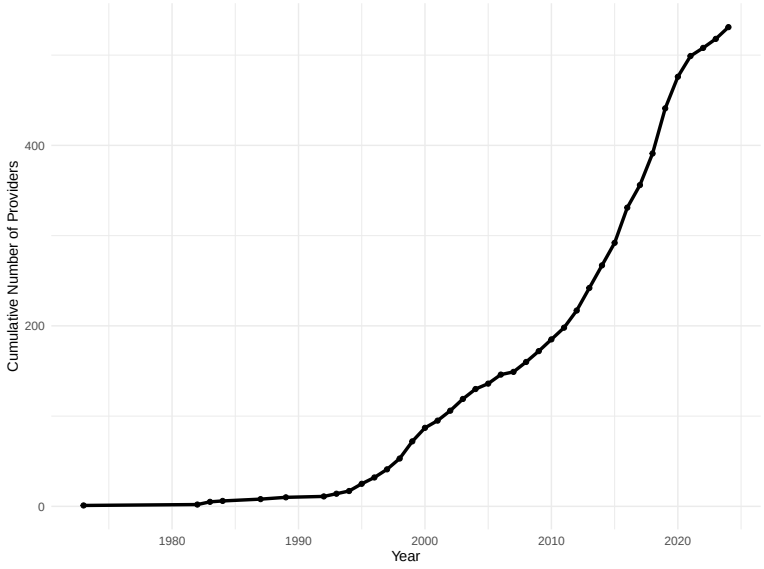
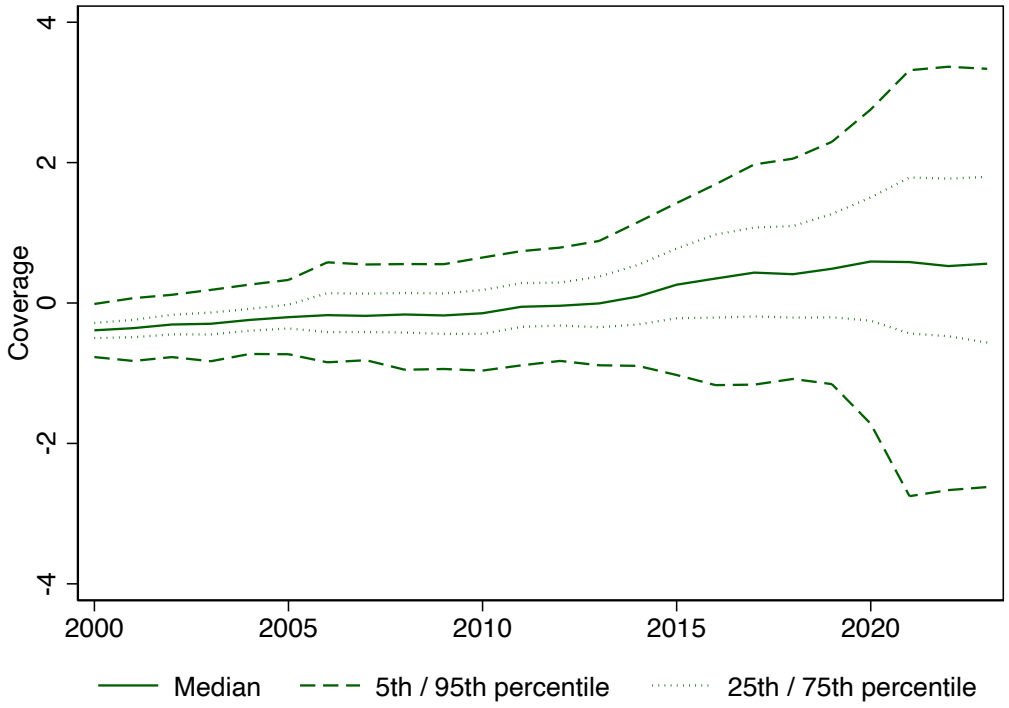


Figure II: Evolution of Distribution of Alternative Data Coverage

This figure illustrates the evolution of firms' exposure to commercially available alternative data. The coverage measure is constructed by summing the cosine similarities between a firm's annual business description (Item 1 of the 10-K filing) and all commercial alternative datasets available in a given year. To facilitate interpretation, we standardize the raw coverage measure over the full sample to have zero mean and unit variance. Panel A reports the distribution of the coverage measure over time. Panel B reports the distribution of the standardized coverage measure across Fama–French 12 industries. The boxes represent the interquartile range, the horizontal line denotes the median, and the whiskers extend to the adjacent values. Industries are ranked by their median coverage.

Panel A: Distribution over Time



Panel B: Coverage across Industries

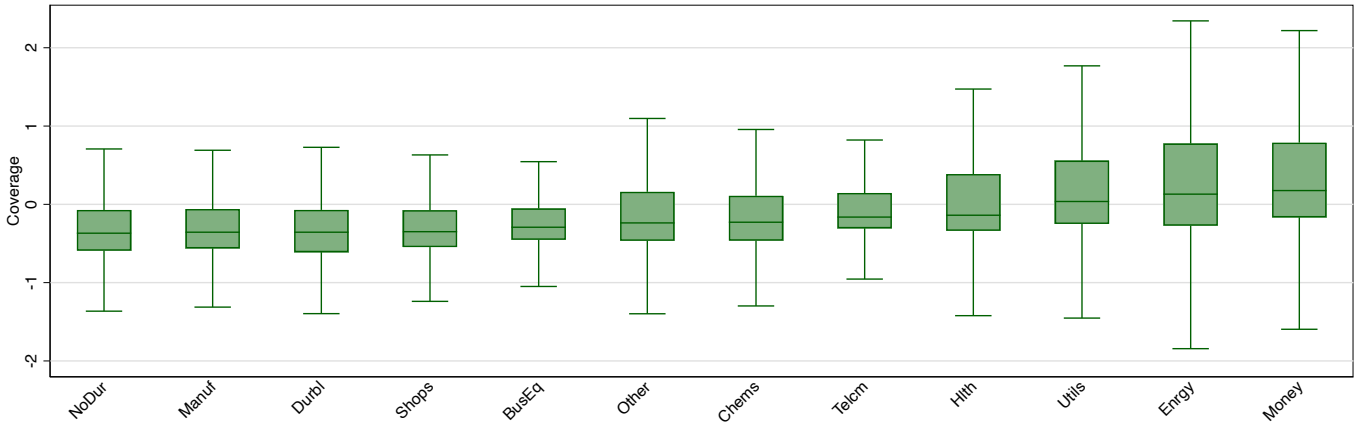


Figure III: Alternative Data Coverage and Hedge Fund Portfolio Excess Weight

This figure reports the dynamic effect of alternative data coverage on hedge funds' portfolio allocations. The specification corresponds to the event-study equation (5). Each circle reports the coefficient on the interaction between $HedgeFund$, $Covered$, and an indicator for a specific quarter relative to the firm's entry into the highest tercile of alternative data coverage. The omitted quarter is the quarter immediately preceding coverage. Dashed lines represent 95% confidence intervals. The dependent variable is the stock's excess portfolio weight relative to its market weight. All specifications include company × year-quarter fixed effects and investor × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

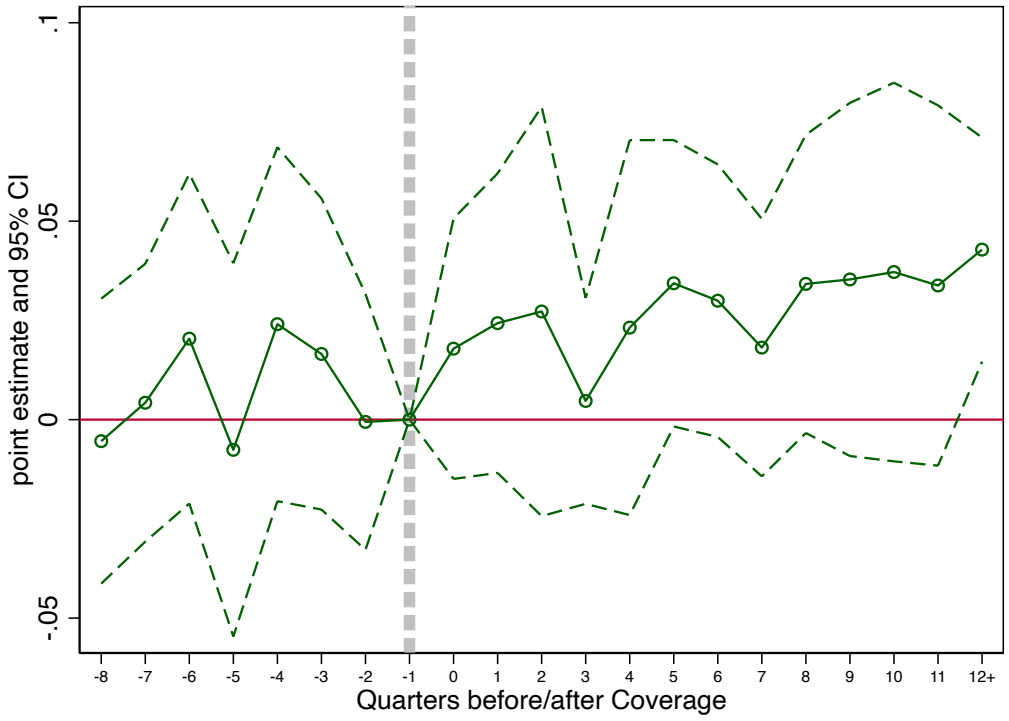
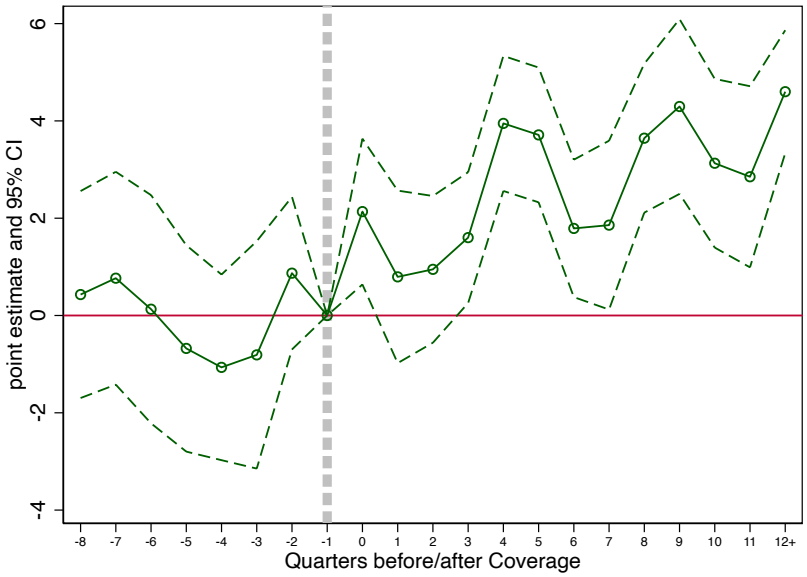


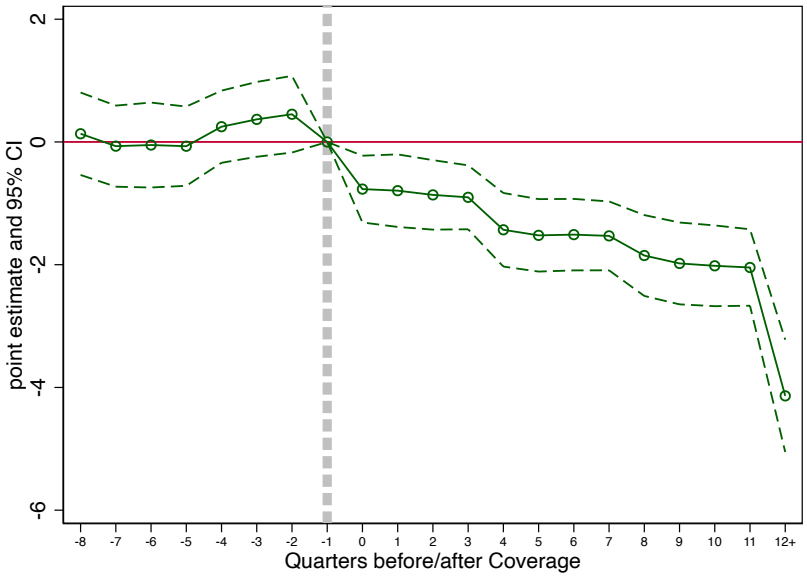
Figure IV: Alternative Data Coverage and Hedge Fund Turnover

This figure reports the dynamic effect of alternative data coverage on hedge funds' turnover and holding duration. The specification corresponds to the event-study equation (5). Each circle reports the coefficient on the interaction between $HedgeFund$, $Covered$, and an indicator for a specific quarter relative to the firm's entry into the highest tercile of alternative data coverage. The omitted quarter is the quarter immediately preceding coverage. Dashed lines represent 95% confidence intervals. In Panel A, the dependent variable is the investor–stock–quarter dollar turnover, defined as the absolute change in the market value of the investor's position, net of passive price appreciation, scaled by the average position size over the current and previous quarters. In Panel B, the dependent variable is the investor–stock–quarter holding duration, defined as the number of consecutive quarters during which the investor has continuously held a positive position in the stock. The measure is reset to one when a new position is initiated and to zero when the position is fully liquidated. All specifications include company × year-quarter fixed effects and investor × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

Panel A: Hedge Fund Turnover



Panel B: Hedge Fund Holding Duration



Internet Appendix

A Alternative Data Marketplaces

Datarade is one of the largest online marketplaces for commercial alternative data and is widely recognized within the industry as a major platform for discovering and sourcing alternative datasets.[23] Its broad coverage of data providers and dataset categories makes it well suited for constructing a comprehensive measure of firms' exposure to commercially available alternative data.[24]

Table A.1 compares Datarade with several major data marketplaces using the DaDaDa dataset of Sun et al. (2026). Datarade lists the largest number of datasets among the set of platforms and offers broad geographic coverage across providers. In addition, it is the only platform in the sample that provides free data samples, which facilitate the inspection of dataset contents. Interestingly, some competing platforms, such as Databricks, directly reference Datarade samples on their marketplace pages, illustrating Datarade's central role in the commercial alternative data ecosystem.

[23] See, for example, https://www.monda.ai/blog/best-data-marketplaces-guide.
[24] See, for example, the review at https://pipeline.zoominfo.com/sales/datarade-review.

Table A.1: Comparison of Major Alternative Data Marketplaces

This table compares major commercial alternative data marketplaces. The statistics are constructed from the DaDaDa dataset of Sun et al. (2026), which aggregates metadata for 16,147 data products listed on nine major marketplaces worldwide, excluding the "Gaming Data" (predominantly horse racing) and "Other" categories (data handling), which together account for less than 1% of the sample and are unlikely to be relevant for financial investors. For each platform, the table reports the number of datasets, the number of distinct data providers, the number of countries represented by at least 5% of listed datasets, and the number of datasets for which a free sample is available.

Platform	# Datasets	# Providers	# Countries	# Free Samples
Datarade	**4392**	**584**	**260**	**255**
AWS Data Exchange	3693	463	14	0
Snowflake	2520	612	260	0
Databricks	1976	203	1	0
Zhejiang Big Data Exchange	838	1	1	0
Canton Data Exchange	786	154	1	0
Beijing International Data Exchange	712	86	1	0
Guiyang Global Big Data Exchange	668	79	1	0
Shanghai Data Exchange	446	73	260	0

Table A.2: Update Frequency of Commercial Alternative Datasets

This table reports the distribution of update frequencies for commercial alternative datasets listed on Datarade. The statistics are constructed using the DaDaDa dataset of Sun et al. (2026). We restrict the sample to Datarade and exclude the "Gaming Data" (predominantly horse racing) and "Other" categories (data handling), which together account for less than 1% of the sample and are unlikely to be relevant for financial investors. The table reports the number of datasets, the corresponding percentage of datasets, and the cumulative percentage for each reported update frequency.

Update Frequency	# Datasets	% Datasets	Cum. % Datasets
Real-Time	986	22.45	22.45
Secondly	71	1.62	24.07
Minutely	66	1.50	25.57
Hourly	118	2.69	28.26
Daily	604	13.75	42.01
Weekly	303	6.90	48.91
Monthly	537	12.23	61.13
Quarterly	90	2.05	63.18
Yearly	88	2.00	65.19
No Update	1128	25.68	90.87
On Demand	401	9.13	100.00

B Examples of Similarity Between Firms and Datasets

Dataset Description

Annual production volumes data streaming service for every refinery, gas processing facility, fractionator, storage and terminals handling gas-liquids in Canada & US. The LPG Almanac Virtual Data Room by Sulpetro Inc. is an interactive data streaming experience that allows our users to quickly visualize key Liquefied Petroleum Gas market regions by basin and filter, aggregate, and regionalize infrastructure. Basic data plan includes 4 years' historical production volume data however, access up-to 20 years (additional costs apply) of annual production volumes data available for market analysis, competitor analysis, and infrastructure projects economics.

LPG Energy Production Volumes US & Canada / Industry Geospatial Production Data Streaming Service Dataset
Sulpetro

High Similarity Firm ($Similarity = 0.72$)

Resources was incorporated in the state of Virginia on July 31, 1998, for the primary purpose of becoming the holding company for Roanoke Gas and its subsidiaries. Effective July 1, 1999, Roanoke Gas and its subsidiaries were reorganized into the holding company structure. Resources is currently composed of the following subsidiaries: Roanoke Gas, Diversified Energy and Midstream. Roanoke Gas, originally established in 1883, was organized as a public service corporation under the laws of the Commonwealth of Virginia in 1912. **The principal service of Roanoke Gas is the distribution and sale of natural gas to residential, commercial and industrial customers within its service territory in Roanoke, Virginia and the surrounding localities.**

Form 10-K, FY 2018
RGC RESOURCES INC

Low Similarity Firm ($Similarity = 0.45$)

lululemon athletica inc. is principally a designer, distributor, and retailer of healthy lifestyle inspired athletic apparel and accessories. We have a vision to be the experiential brand that ignites a community of people through sweat, grow, and connect, which we call "living the sweatlife." Since our inception, we have fostered a distinctive corporate culture; we promote a set of core values in our business which include taking personal responsibility, nurturing entrepreneurial spirit, acting with honesty and courage, valuing connection, and choosing to have fun.

Form 10-K, FY 2018
lululemon athletica inc.

Dataset Description

Q - How is the data generally sourced? A - The ratings are generated using data from bank Call Reports filed with the FDIC quarterly. The Call Reports contain vast amounts of data on the financial and operating results of each bank and provides a robust and consistent set of measures from which to generate Efficient Frontiers. **Because all banks with FDIC insurance must file Call Reports with the FDIC, we can provide ratings on more than 99% of all FDIC insured banks.**

U.S. Bank Operational Efficiency & Business Quality Ratings Dataset
Hoeg & Company

High Similarity Firm ($Similarity = 0.62$)

Summit Financial Group, Inc. Company or Summit is a 1.39 billion financial holding company headquartered in Moorefield, West Virginia. We provide community banking services primarily in the Eastern Panhandle and South Central regions of West Virginia and the Northern region of Virginia. We provide these services through our community bank subsidiary, Summit Community Bank Summit Community or Bank). We provide a wide range of community banking services, including demand, savings and time deposits; commercial, real estate and consumer loans; letters of credit; and cash management services. **The deposits of Summit Community are insured by the Federal Deposit Insurance Corporation ("FDIC").**

Form 10-K, FY 2013
Summit Financial Group

Low Similarity Firm ($Similarity = 0.46$)

Empire Global Gaming, Inc. (the Company, We or Our) was incorporated under the laws of the State of Nevada on May 11, 2010 to participate in the worldwide gaming market. We offer custom game manufacturing, gaming solutions, supplies and custom game marketing. We provide consulting and advisory services to the gaming industry.

Form 10-K, FY 2013
Empire Global Gaming, Inc.

C Examples of Datasets Triggering Firms' Treatment

Firm Description (Treated in 2021)

MidAmerican Energy, an indirect wholly owned subsidiary of BHE, is a United States regulated electric and natural gas utility company headquartered in Iowa that serves 0.8 million retail electric customers in portions of Iowa, Illinois and South Dakota and 0.8 million retail and transportation natural gas customers in portions of Iowa, South Dakota, Illinois and Nebraska. [...] **Variations in weather, economic conditions and various conservation, energy efficiency and private generation measures and programs can impact customer usage.**

Form 10-K, FY 2020
MidAmerican Energy Co.

Triggering Dataset

Hyperlocal renewables and site-specific power plant production modelling - Accurate grid forecasting - Demand forecasts for power, gas and heat suppliers - Risk and uncertainty forecasting based on numerous models and ensemble model runs **- Weather and climate analysis for assessment of new assets** - Real time data and global forecasts for up to 30 days ahead.

Energy modelling - solar power and radiation + wind power forecasts
Meteosource Weather API

Firm Description (Treated in 2022)

We are a global media and technology company with three primary businesses: Comcast Cable, NBCUniversal and Sky. We were incorporated under the laws of Pennsylvania in December 2001. Through our predecessors, we have developed, managed and operated cable systems since 1963. Through transactions in 2011 and 2013, we acquired NBCUniversal, and in 2018, we acquired Sky. We present our operations in five reportable business segments: (1) Comcast Cable in one reportable business segment, referred to as Cable Communications; (2) NBCUniversal in three reportable business segments: Media, Studios and Theme Parks (collectively, the NBCUniversal segments and (3) Sky in one reportable business segment.

Form 10-K, FY 2021
Comcast Corporation

Triggering Dataset

Using our intelligently designed data dashboard, you can quickly understand how Comcast Corporation (CMCSA) is lobbying the U.S. government, how much they're spending on it, and most importantly - the bills and specific issues on which they lobby. Gain an informational edge over the market with our Lobbying Data Intelligence Platform. [...] Perform analysis by company, lobbyist, lobbying firm, government agency, or issue. For lobbying firms: understand your competitors. Understand who is registering with who. Gain insight on quarterly reports and specific issues other firms are lobbying on.

Comcast Corporation (CMCSA) all U.S. Lobbying: all historical lobbying contracts, government bills & agencies, and critical issues lobbied on.
LobbyingData.com

D Tables

Table D.3: Alternative Data Coverage and Quantitative Hedge Funds

This table examines whether the response of hedge funds to alternative data coverage differs for quantitative hedge funds. Regressions are estimated at the investor–stock–quarter level. The dependent variables are excess portfolio weight relative to the stock's market weight in Column (1), dollar turnover in Column (2), an indicator for portfolio entry or exit (Churn) in Column (3), and holding duration in quarters in Column (4). The variable $Covered \times Post$ is an indicator equal to one after a stock enters the highest tercile of alternative data coverage. The coefficient on $Covered \times Post \times Hedge\ Fund$ measures the overall response of hedge funds relative to other investors. The coefficient on $Covered \times Post \times Hedge\ Fund \times Quant$ captures the additional response of quantitative hedge funds relative to other hedge funds. Section IV.D describes our classification of quantitative hedge funds. The sample covers the period from 2012 to 2023. All specifications include company × year-quarter fixed effects and investor × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

	Excess Weight (%)	Dollar Turnover (%)	Churn	Holding Duration
	(1)	(2)	(3)	(4)
Covered × Post × Hedge Fund	0.055***	2.039***	0.004**	-2.097***
	(0.018)	(0.477)	(0.002)	(0.388)
Covered × Post × Hedge Fund × Quant	-0.024	1.110***	0.005*	-0.720***
	(0.017)	(0.376)	(0.002)	(0.262)
Company × Year-Quarter FE	Yes	Yes	Yes	Yes
Investor × Year-Quarter FE	Yes	Yes	Yes	Yes
Observations	1.77e+07	1.77e+07	1.77e+07	1.77e+07
R^2	0.58	0.41	0.37	0.53

Table D.4: Alternative Data Coverage and High Hedge Fund Ownership

This table reports the effect of alternative data coverage on the likelihood of high hedge fund ownership. Regressions are estimated at the firm–quarter level using a stacked difference-in-differences design. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variable is an indicator equal to one if hedge funds own more than 0%, 1%, 2%, 3%, 4%, 5%, and 10% of the firm's equity in Columns (1)–(7), respectively. The variable *Covered* × *Post* is an indicator equal to one for treated firms after they enter the highest tercile of alternative data coverage. All specifications include company × cohort fixed effects, year-quarter × cohort fixed effects, and control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns. Standard errors are two-way clustered by company and year-quarter.

	Share > 0%	Share > 1%	Share > 2%	Share > 3%	Share > 4%	Share > 5%	Share > 10%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Covered × Post	-0.005	0.002	0.023***	0.024***	0.023***	0.021**	0.015*
	(0.004)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.007)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.84	0.79	0.78	0.78	0.79	0.80	0.80

Table D.5: Hedge Fund Ownership and Stock Return Comovement

This table reports the association between hedge fund ownership and stock return comovement. Regressions are estimated at the firm–quarter level using the stacked difference-in-differences sample. For each treated firm, we identify the quarter in which it first enters the highest tercile of alternative data coverage and match it to five control firms based on market capitalization and pre-treatment ownership composition across investor types. The sample is then stacked across treatment cohorts. The dependent variables are quarterly return comovement measures. Specifically, return comovement is measured as the R^2 from a daily regression of a firm's stock returns on the contemporaneous, lagged, and lead equal-weighted returns of all other firms in the reference group. The reference group is the entire market in Columns (1) and (2), firms in the same industry in Columns (3) and (4), and firms in the same alternative-data similarity tercile in Columns (5) and (6). Even-numbered columns report the logit transformation of the corresponding R^2 measure. The main independent variable is the ownership share held by hedge funds. All specifications control for log market capitalization, market beta, leverage, dividend payments, log book equity, and lagged one-quarter stock returns, and include company × cohort fixed effects and year-quarter × cohort fixed effects. Standard errors are two-way clustered by company and year-quarter.

	$R^{2,mkt}$	$Logit(R^{2,mkt})$	$R^{2,ind}$	$Logit(R^{2,ind})$	$R^{2,data}$	$Logit(R^{2,data})$
	(1)	(2)	(3)	(4)	(5)	(6)
HF Ownership	-0.002***	-0.015***	-0.002***	-0.016***	-0.002***	-0.015***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.000)	(0.002)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter × Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99,577	99,577	99,577	99,577	99,577	99,577
R^2	0.77	0.75	0.80	0.78	0.77	0.75

E Figures

Figure E.1: Alternative Data Coverage and Hedge Fund Churn

This figure reports the dynamic effect of alternative data coverage on hedge funds' turnover and holding duration. The specification corresponds to the event-study equation (5). Each circle reports the coefficient on the interaction between $HedgeFund$, $Covered$, and an indicator for a specific quarter relative to the firm's entry into the highest tercile of alternative data coverage. The omitted quarter is the quarter immediately preceding coverage. Dashed lines represent 95% confidence intervals. The dependent variable is the investor–stock–quarter churn indicator, equal to one if the investor either initiates or completely liquidates a position in the stock during the quarter, and zero otherwise. All specifications include company × year-quarter fixed effects and investor × year-quarter fixed effects. Standard errors are three-way clustered by investor, company, and year-quarter.

