# Data Scientists on Wall Street[*]

Ling Cen[†] Bing Han[‡] Yanru Han[§] Chanik Jo[¶]

December 14, 2024

**Abstract**

Financial institutions have significantly increased their recruitment of data scientists in the last two decades. We find that the number of data scientists employed by financial institutions causally affects their ability to earn abnormal profits. Data scientists' ability to generate abnormal profits on a stock is positively correlated to the concentration of data scientists across all institutional investors holding the stock. Institutional investors strategically adjust portfolio allocation and recruitment decisions to maximize the benefits generated by their data scientists. Consistent with the notion that the competition among data scientists speeds up the production and trade of private information, we also show that the concentration of data scientists covering a stock reduces its price informativeness in the capital market.

***JEL Classification:*** *G10, G11, G14, G23*

***Keywords:*** *Data Scientists, Information Competition, Price Informativeness, Institutional Investors*

[*]We are grateful to Lauren Cohen, Michael Hertzel, Clemens Sialm, Yan Xiong, and Liyan Yang as well as seminar participants at Yonsei University for insightful comments and suggestions. All errors are our own. First draft: May 1, 2024

[†]The Chinese University of Hong Kong, Email: ling.cen@cuhk.edu.hk

[‡]The University of Toronto and The Chinese University of Hong Kong, Email: binghan@cuhk.edu.hk

[§]Stevens Institute of Technology, Email: yhan47@stevens.edu

[¶]The Chinese University of Hong Kong, Email: chanikjo@cuhk.edu.hk

# 1 Introduction

The "big data" revolution has fundamentally reshaped how institutional investors acquire and process information in the capital market (Goldstein et al., 2021). In addition to relying on standardized data from corporate disclosures (e.g., annual reports and corporate announcements) and information intermediaries (e.g., Bloomberg, Reuters, and financial analysts), financial institutions attempt to beat the market by recruiting data scientists who collect, maintain, and analyze unstructured alternative data with artificial intelligence techniques.[1] Our analysis based on personal profiles from the Revelio Lab database suggests that the total number of data scientists employed by financial institutions has more than quadrupled (i.e., 11,799 to 57,050) from 2008 to 2021. With the fast-growing demand for skilled data scientists outpacing the growth of talent supply, financial institutions are competing not only among themselves but also against technology companies to attract the best workers in data science.[2]

Our paper explores three sets of research questions related to the increasingly important role of data scientists in the capital market. First, why do financial institutions aggressively recruit data scientists? Do data scientists help financial institutions identify abnormal returns, i.e., alphas, in portfolio management? Second, if data scientists do generate significant benefits, do financial institutions strategically adjust portfolio allocation decisions in the financial market and recruitment decisions in the labor market to maximize these benefits?

[1]For example, in May 2017, JP Morgan released a report on "Big Data and AI Strategies", which includes 594 alternative data sources that institutional investors can extract useful information for trading.

[2]The Data & AI Human Capital Report, issued by Darwin X, shows that the total data talent pool constitutes only 1.4% of the total workforce of major banks. The data and AI workforce is expected to grow by 17% in 2024 and 38% in 2025. While all financial institutions are competing for data scientists, the report suggests that the gap between leaders and laggards is widening.

Last but not least, we are interested in the aggregated impacts of data scientist recruitment on the capital market. In particular, we examine whether the total number and the distributional concentration of data scientists covering a stock affect its price informativeness.

To answer our research questions empirically, we link institutional investors' employment of data scientists to their stock holdings. We collect detailed resumes of employees hired by institutional investors from the Revelio Lab database, which gathers career history data from various unstructured public online sources, including LinkedIn. The initiation of our sample in 2008 coincides with the proliferation of online career websites during the early 2000s, with the career data provided by the Revelio Lab achieving representativeness towards the latter part of the same decade. After merging the Revelio data with institutional holding data from the Refinitiv/Thomson Reuters Global Ownership database, we start with a merged dataset containing records of 3,265,145 unique workers affiliated with 7,588 distinct institutional investors from 2008 to 2021. We then identify data scientists within this group based on the ONET occupation codes.[3] This approach allows us to identify 326,627 unique data scientists employed by 3,126 institutional investors in our sample period. Based on the main tasks and required skills from job descriptions, we classify data scientists into three groups – data collectors, maintainers, and analyzers.

Our analysis shows that institutional investors who hire more data scientists achieve higher trading profitability, which is consistent with the notion that data scientists help institutions identify mispriced assets. We find that, on average, each additional data scientist hired by an investor leads to a 0.004 percentage point increase per quarter in CAPM alpha,

[3]The specific ONET codes corresponding to data scientist roles are listed in Appendix Table A2.

translating to a 13% improvement over the average trading profitability in our sample. The results are robust across different measures of trading profitability and various roles of data scientists (i.e., data collectors, maintainers, and analyzers). Among these, data analyzers stand out as the most impactful category, underscoring the critical role of data analysis in generating investment insights. The above correlation between the number of data scientists and institutional performance might be driven by omitted variables that affect both. To address this endogeneity concern, we leverage an instrumental variable approach using the introduction of data science undergraduate programs as an exogenous shock to the future (i.e., four-year future) local supply of data scientists. We show that the increase in the recruitment of data scientists, as an outcome of exogenous increases in local labor supply, leads to significant improvements in trading profitability, validating the causal link between data scientist recruitment and institutional investors' ability to "seek alpha".

We further explore the conditions under which investors gain the most from hiring data scientists and we focus on a unique dimension of data scientist concentration, defined as the Herfindahl-Hirschman Index (HHI) of data scientists across all institutions holding a stock. For example, think about a stock held by ten institutional investors. In the first case, one institutional investor has ten data scientists and the other nine have none (HHI=1). In the second situation, each institutional investor has one data scientist (HHI=0.1).[4] While the total number of data scientists covering this stock is the same under these two situations, the first case obviously has a much higher data scientist concentration relative to the second case. When data scientist concentration is higher, the informational advantage of institutions

[4]This is from $\sum_{i=1}^{10}(1/10)^2 = 0.1$.

that employ more data scientists could be more pronounced. Conversely, if each investor holding a stock employs a similar number of data scientists, the competition to uncover and act on useful information becomes more intense, diminishing this informational advantage (e.g., Holden and Subrahmanyam, 1992; Xiong et al., 2024). Our analysis reveals that the positive impact of hiring data scientists on trading profitability is indeed amplified when data scientist concentration is higher.

Our findings indicate that institutional investors strategically adjust their portfolio allocation and recruitment decisions to maximize the benefits provided by data scientists. First, we show that investors who hire more data scientists tend to concentrate their holdings on a smaller set of stocks. This strategic focus allows their data scientists to generate information advantage, thereby improving trading performances in a competitive environment. Moreover, we find that investors strategically tilt their asset allocation toward stocks with higher data scientist concentration, where their data scientist employment allows them to enjoy an information monopoly advantage, and subsequently boost their trading profitability.

Beyond portfolio decisions, we find that investors actively respond to competitive pressures in the labor market. Investors appear to keep up with their rivals' hiring activities and seek to close gaps in data scientist staffing. When investors find themselves lagging behind leading competitors in hiring data scientists, they respond by increasing their recruitment efforts. This behavior reflects a talent race within the finance industry, where investors strive to maintain a competitive edge by continually expanding their data science capabilities. These strategic moves in both the financial market and the labor market underscore the importance that investors place on the unique value offered by data scientists.

Our findings so far reveal that while investors gain an information advantage from employing data scientists, this advantage can come at the expense of market efficiency. Specifically, we demonstrate that the concentration of data scientists among a few investors diminishes price informativeness. A one standard deviation increase in data scientist concentration leads to an 11% decrease in price informativeness, highlighting the efficiency cost of concentrated data scientist coverage. When data scientists are concentrated among a few institutional investors, the "information monopoly" of investors with more data scientists gives them a stronger incentive to delay the incorporation of their private information into stock prices, resulting in less efficient price formation.

To strengthen our causal inference, we take advantage of mergers and acquisitions (M&As) among institutional investors as a source of exogenous variation in data scientist concentration. We argue that M&A decisions are generally not motivated by the specific data scientist resources of the acquirer and target institutions. Furthermore, a merger between two investors holding different stocks can lead to divergent changes in data scientist concentration, depending on the pre-merger conditions of each stock. This heterogeneity provides a quasi-experimental setting to examine the impact of data scientist concentration on price informativeness. Our analysis shows that exogenous decreases in data scientist concentration following financial institution mergers lead to improved price informativeness, further supporting our hypothesis that concentrated data scientist coverage negatively affects market efficiency.

Our paper is closely related to existing studies on the role of big data in the financial market. Farboodi et al. (2022) develop a quantitative measure of data usage by investors.

Farboodi et al. (2019) and Veldkamp (2023) provide a theoretical framework and a set of tools to value data as an asset. Other studies in this field have investigated a large number of implications that big data has generated in the capital market, including price informativeness (Dugast and Foucault, 2018), market efficiency (Martin and Nagel, 2022), capital allocation (Dugast and Foucault, 2024), forecast accuracy (Chi et al., 2024), forecast horizon (Dessaint et al., 2024), endogenous data skill acquisition (Huang et al., 2022), and disciplinary effect of corporate managers (Zhu, 2019). Unlike data that can be shared with a positive externality, we focus on data scientists who cannot be employed by multiple institutional investors at the same time. While institutional investors compete for data talent, we find the distributional concentration of data scientists across institutional investors affects price informativeness.

Our paper also contributes to the growing literature on the labor market of financial workers. Most studies in this field focus on the finance wage premium, i.e., compensation in the finance industry has been much higher than that in other sectors since the 1990s (e.g., Philippon and Reshef, 2012; Boustanifar et al., 2017; Célérier and Vallée, 2019). Due to data limitations, only a very small number of studies have investigated the determinants of career choices and progressions in buy-side financial institutions. For example, Oyer (2008) finds that stock market conditions have a large effect on whether MBA graduates go directly to Wall Street after graduation. Ellul et al. (2019) show that managers working for liquidated funds due to poor performance suffer demotion or compensation loss after job turnovers. We identify a prevailing trend of recruiting data scientists among institutional investors and show that data scientists causally help financial institutions generate abnormal returns.

The rest of this paper is organized as follows. Section 2 describes the data and variable construction. Section 3 presents empirical results addressing three research questions. Section 4 concludes.

## 2 Data Sources and Summary Statistics

### 2.1 Data Scientist Measures

We obtained detailed resumes of employees hired by institutional investors from the Revelio Lab database, which collects data from various unstructured public career websites, including LinkedIn. This dataset provides detailed information related to employment histories, including job titles, skill requirements, employer (company) names, employment duration, and descriptions of job responsibilities. The platform began in the early 2000s, and for consistency with prior research (e.g., Liang et al., 2023; Cai et al., 2024), our sample covers the period from 2008 to 2021. During this period, Revelio has documented 646,940,525 unique employment experiences from 220,358,527 individuals worldwide, associated with 17,163,699 different public and private companies.

We pair employers in the Revelio data with institutional investors from the Thomson Reuters Global Ownership database using a name-matching algorithm. We then manually verified these matches to ensure their accuracy. The combined Revelio-Global Ownership dataset includes records of 2,908,292 unique workers connected to 7,408 distinct institutional investors, covering a sample period from 2008 to 2021.

Data scientists are identified based on the occupation codes (ONET codes) associated with their positions. By reviewing the tasks described in the ONET occupation database, we

partition data scientists into three groups according to their roles and functions: data collection, data analytics, and data maintenance. Each role has clearly defined responsibilities: “data collection" refers to the tasks of gathering and organizing data, “data analytics" refers to the tasks of analyzing data for making business decisions, and “data maintenance" refers to the tasks of storing and protecting data by maintaining proper hardware. The specific ONET codes for these data scientist roles are listed in Appendix Table A2. Our final dataset includes 124,947 unique data scientists employed by 1,957 different investors.

Our dataset on data scientists is constructed based on online personal profiles, instead of corporate reports released by firms. This approach mitigates the concerns about corporate-disclosure-based data that companies may have strategic incentives to disclose (or not to disclose) the recruitment of data scientists. Previous studies using the same dataset find that online profiles are more likely to represent white-collar workers than blue-collar workers (Li et al., 2022) and workers with online profiles have a higher level of computer literacy than those without online profiles. This limitation is minor for our study since data scientists are typically white-collar workers with a high level of computer literacy. They are well-represented in online profiles.

While it is straightforward to count the number of data scientists employed by a financial institution, computing the number of data scientists covering a stock is nontrivial. Obviously, for a single stock, a data scientist of a major shareholder is not equivalent to a data scientist of a minor shareholder. Following this intuition, we compute the ownership-weighted average number of data scientists covering a specific stock as follows:

$$DS\ Coverage_{j,t} = \sum_{i} NumDS_{i,t} \times \frac{Shares\ Held_{i,j,t}}{Shares\ Outstanding_{j,t}}, \tag{1}$$

where $j$ denotes a stock, $i$ denotes an investor, and $t$ denotes a time period. $NumDS_{i,t}$ is the number of data scientists employed by investor $i$ at time $t$. $Shares\ Held_{i,j,t}$ represents the number of shares of stock $j$ held by investor $i$ at time $t$. $Shares\ Outstanding_{j,t}$ is the total number of shares outstanding of stock $j$ at time $t$. The ratio of the two (i.e., $Shares\ Held_{i,j,t}$ scaled by $Shares\ Outstanding_{j,t}$) is the percentage of all outstanding shares of stock $j$ held by investor $i$ at time $t$. By constructing this measure, we implicitly assume that data scientists of major shareholders exert a higher influence on stock prices relative to those of minor shareholders or those of financial institutions that have no position in this stock.

Another key variable of interest, $DS\ HHI$, measures the concentration of data scientists for a specific stock. Our intuition is captured by the following example. Consider two stocks, Stocks 1 and 2, and four institutional investors, Investors $a$, $b$, $c$, and $d$. Investors $a$ and $b$ each hold 50% of Stock 1, while Investors $c$ and $d$ each hold 50% of Stock 2. Investors $a$, $b$, $c$, and $d$ employ 10, 0, 5, and 5 data scientists, respectively. Under this assumption, both the total number of data scientists (10) and the ownership-weighted number of data scientists (5) are the same for Stocks 1 and 2. However, the concentration of data scientists differs significantly across the two stocks. Specifically, the data scientists covering Stock 1 are concentrated with Investor $a$, whereas those covering Stock 2 are evenly distributed between Investors $c$ and $d$.

Following this intuition, we compute the concentration of data scientists for a specific stock in a similar spirit as a Herfindahl index, as outlined in the following equation:

$$DS\ HHI_{j,t} = \sum_{i} \left(\frac{NumDS_{i,t} \times \frac{Shares\ Held_{i,j,t}}{Shares\ Outstanding_{j,t}}}{DS\ Coverage_{j,t}}\right)^2, \tag{2}$$

where $j$ denotes a stock, $i$ denotes an institutional investor, and $t$ denotes a time period.

By construction, $DS\ HHI$ is a stock-level variable that is strictly greater than 0 and less than or equal to 1 (i.e, $DS\ HHI_{j,t} \in (0, 1]$), where a value of 1 indicates the highest level of concentration (only one institutional investor hires data scientists) and a value close to zero indicates a more diversified hiring of data scientists across multiple investors. From an individual firm's perspective, data scientist concentration depends on the employment of data scientists of other financial institutions holding the same stock. More importantly, when one financial institution recruits more data scientists, it is likely that some stocks in its holding experience an increase in data scientist concentration and others experience opposite changes.

### 2.1.1 Other Data Sources

The quarterly 13F institutional investors' holdings data are retrieved from Refinitiv/Thomson Reuters Global Ownership Database. The dataset reports holdings of any institutions with assets under management of $100 million or more, covering a wide range of financial institutions, including mutual funds, hedge funds, bank trusts, pension funds, insurance companies, and sovereign wealth funds.

In addition, stock price information is retrieved from CRSP, and corporate performance data are obtained from Compustat. In our identification strategy relying on the mergers and acquisitions of financial institutions, the M&A information is retrieved from the SDC M&A database.

### 2.2 Summary Statistics

In Appendix Table A3, we present the top 50 institutional investors ranked by their data scientist employment in 2021. The top three institutional investors are Morgan Stanley, Credit Suisse, and Goldman Sachs. The list also includes many prominent hedge funds, such as BlackRock Alternatives Management and AQR Capital Management.

**[Insert Figures 1 and 2 Here]**

In Figure 1, we plot the time series for the total number of data scientists hired by all institutional investors with an annual frequency. The figure shows a clear pattern that the financial institutions' employment of data scientists significantly increases during our sample period. Specifically, there are, in total, 11,799 data scientists employed by institutional investors in 2008. This number reaches 57,050 in 2021, which is more than quadrupled relative to the 2008 level.[5] We also observe that the speed of employing data scientists does not slow down after 2020.

Figure 2 provides the time-series patterns for the mean of data scientist concentration across all stocks. We observe that the data scientist concentration dropped after the subprime crisis and, since then, the data scientist concentration has been gradually increasing. There are two possible (and mutually non-exclusive) reasons driving this pattern. One possible reason is that some major financial institutions recruit more data scientists than other

[5]It is possible that this trend is partially contributed by the fact that the online profile for 2021 is more complete than that for 2008. To address this concern, we also verify this pattern by using the Lightcast online job post data and we find the same pattern.

financial institutions. Another possibility is that financial institutions make strategic adjustments to their holdings so that they are more likely to invest in stocks where they have advantages in information processing ability (e.g., having relatively more data scientists than other investors). We will test both conjectures in the following sections.

**[Insert Table 1 Here]**

Table 1 provides summary statistics of all key variables. Panel A reports institutional investor-level statistics, while firm-level statistics are reported in Panel B. On average, institutional investors in our sample hire two data scientists, though the distribution is highly skewed, with over 75% of investors not employing any data scientists during the sample period. Among the three categories of data scientists, those focused on data analytics are the most numerous. The average investor has a natural logarithm of total assets under management of 19.978.

The holding-weighted average number of data scientists covering firms in our sample is 17. Consistent with the investor-level summary statistics, data scientists focused on analytics make up the largest proportion of data scientist coverage for firms. The average concentration of data scientists is 0.409 in our sample.

**[Insert Table 2 Here]**

We first run panel regressions to identify the type of financial institutions that are more likely to recruit data scientists. The dataset for these tests is organized at the institution-year level. The dependent variable in Panel A is the total number of data scientists employed

by a specific institutional investor in year $t$. We include common institutional characteristics as independent variables, including asset size, asset turnovers, the number of industries covered in portfolios, institution types, and the number of data-science-related undergraduate programs of local universities. In column (2), we control for time fixed effects to eliminate the impact of market-wide shocks. In column (3), we control for institution type×time fixed effects to eliminate the market-wide shocks specific to each type of financial institution. Results in Panel A of Table 2 suggest that financial institutions that have high turnovers in investment, cover more industries in portfolios, and have larger assets under management are likely to recruit more data scientists.[6] Relative to financial institutions classified as financial advisors, hedge funds employ more data scientists while pension funds and banks employ fewer. The most interesting result in Panel A of Table 2 is that the number of data scientists recruited by financial institutions is positively correlated with the number of data-science-related undergraduate programs offered by local universities, suggesting that the supply of data scientists in the local labor market is likely to affect financial institutions' employment of data scientists in equilibrium. Our identification strategy that relies on the establishment of new data science undergraduate programs in local universities in Section 3.1 is motivated by this important correlation in Panel A of Table 2.

In addition, we also examine which types of stocks are more likely to be covered by data scientists at the stock-year level. Panel B of Table 2 presents the cross-sectional correlation between the number of data scientists covering each firm and firm characteristics. The dataset for these tests is organized at the firm-year level. Our results suggest that the

[6]Since $LogTNA$ and $Num\ NonDS\ Employee$ are highly correlated, we show in unreported results that investors with higher assets under management are more likely to hire data scientists when $Num\ NonDS\ Employee$ is excluded from the regression.

ownership-weighted number of data scientists employed by financial institutions investing in a stock is positively correlated with firm size, institutional ownership, leverage, cash holding, and the existence of reported financial fraud, but is negatively correlated with institutional ownership concentration, Tobin's Q, and firm age.

# 3 Empirical Results

## 3.1 Seeking Alpha: Why Do Financial Institutions Recruit Data Sciencits?

Given the fact that financial institutions are competing for data scientists in the labor market, it is important to understand the benefits or strategic advantages brought by data scientists. For most institutional investors, particularly actively managed funds, the primary goal is to achieve the highest trade-off between return and risk, i.e., maximizing the Sharpe ratio. Data scientists play a crucial role in "connecting the dots" of all public and private information available to financial institutions, generating accurate predictions with superior information processing, and helping firms seek alpha on mispriced assets. In this section, we examine whether the employment of data scientists is positively correlated with trading profitability, i.e., alphas, of their affiliated financial institutions. More importantly, we aim to understand whether this positive correlation is also affected by the distribution of data scientists across all major institutions investing in one specific asset.

To this end, we estimate the following regression for 13F investor $i$ at quarter $t$:

$$Alpha_{i,t+1} = \beta_0 + \beta_1 NumDS_{i,t} + Investor\ Controls + Fixed\ Effects, \tag{3}$$

where $Alpha_{i,t+1}$ denotes trading profitability that we construct to examine whether an institution with data scientists has the ability to change its holdings in the direction of subsequent

abnormal returns (alphas), following Kumar et al. (2020) and Bonelli and Foucault (2023).[7] Specifically, trading profitability is the sum of the portfolio-weight change from $t-1$ to $t$ times alphas of each stock earned from $t$ to $t+1$. In this way, we can estimate the contribution of data scientists to the institution's quarterly alphas through active trading. Alphas are computed based on the CAPM, Fama-French 3-factor, and Fama-French 4-factor models. CAPM $\alpha_{t+1}$, for instance, is calculated as:

$$CAPM\ \alpha_{i,t+1} = \sum_{j}(Weight_{i,j,t} - Weight_{i,j,t-1}) \times CAPM\ \alpha_{j,t+1}, \tag{4}$$

where $Weight_{i,j,t}$ represents the portfolio weight the investor $i$ holds in stock $j$ in quater $t$ and $CAPM\ \alpha_{j,t+1}$ is the CAMP alpha of the stock $j$. The CAMP alpha of the stock $j$ is estimated at a monthly frequency, and we accumulate monthly alphas to compute quarterly alphas, which are then multiplied by 100 to be presented in percentage. A higher CAPM alpha for an investor indicates that they either had increased their holdings in stocks that subsequently generated positive alpha or had reduced their holdings in stocks that ended up earning negative alpha. Thus, the CAPM alpha captures the investor's trading profitability.

The key independent variable is $NumDS_{i,t}$, which is the number of data scientists employed by investor $i$ at time $t$. Investor-level control variables include the log of total assets under management of the investor ($LogTNA$), the log of the number of stocks in the investor's portfolio ($Log\ Number\ of\ Firm$), the investor's portfolio turnover ($Turnover$), a dummy variable that equals to one if the investor's portfolio covers two or fewer industries ($FewIndDummy$), the ownership-weighted average of the log of the market capitalization

[7] Kacperczyk et al. (2014) and Jiang and Zheng (2018) construct similar measures based on the deviation of the fund's holdings relative to a benchmark portfolio to measure the stock picking ability.

of stocks in the portfolio ($Log\ Market\ Cap$), the ownership-weighted average of the trading volumes of stocks in the portfolio ($Volume$), and the ownership-weighted average of the gross profit margin of stocks in the portfolio ($Gross\ Profit$).

We include investor fixed effects to account for time-invariant unobservable characteristics specific to each investor that could influence trading profitability as well as average performance. Additionally, year fixed effects are included to control for common time trends affecting all investors in a given period. In Equation (3), our coefficient of interest is $\beta_1$ that measures the effect of hiring data scientists on investor trading profitability. A positive and significant $\beta_1$ would indicate that institutional investors with more data scientists achieve higher trading profitability.

**[Insert Table 3 Here]**

Panel A of Table 3 presents the estimates from Equation (3). The coefficients on $NumDS$ ($\beta_1$) are positive and significant across all columns, indicating that investors who hire more data scientists exhibit higher trading profitability. The findings are robust across different alpha measures and economically significant. For example, in column (1), each additional data scientist hired by the investor is associated with a 0.004 percentage point increase in trading profitability as measured by CAPM alpha, which represents a 13% improvement over the average trading profitability of investors in our sample.

In Panel B, we divide data scientists into three groups based on their roles and functions: data analytics, data collection, and data maintenance. We then analyze the relationship between investor trading profitability and the hiring of each of these three types of data

scientists. The coefficients of interest in columns (1)-(3) are all positive and significant, indicating that hiring any of these three types of data scientists is positively correlated with investor performance. In column (4), we include all three types of data scientists in the regression. The result suggests that data scientists focused on data analytics exhibit the strongest and most robust positive relationship with investor trading profitability, aligning with the intuition that data analysis provides the most direct and valuable insights for investment decisions.

The relationships observed between the hiring of data scientists and trading profitability thus far reflect correlations, but these correlations do not guarantee causality. Specifically, the positive correlation does not ensure that increasing the number of data scientists will lead to improved trading profitability for investors. Reverse causality is a possibility: investors who are already performing well may have greater financial capacity to hire additional employees, including data scientists. Moreover, unobserved common variables could influence both trading profitability and the decision to hire data scientists simultaneously.

We address endogeneity concerns by exploiting an identification strategy based on an exogenous shock to the local availability of data scientists. This allows us to examine the causal impact of this increased availability on the trading profitability of investors located in the same states. In Panel A of Table 2, we show that the number of data scientists hired by institutional investors is positively correlated with the number of data-science-related undergraduate programs offered by local universities, suggesting that the supply of data scientists in the local labor market is likely to affect institutional investors' employment of data scientists in equilibrium. Our identification strategy hinges on the establishment of new

data science undergraduate programs, which are largely independent of local institutional investors' actions. While one might argue that local labor market demand could influence the establishment of these programs, the four-year lag between the introduction of new programs and the availability of graduates makes it difficult for institutional investors to time or control the supply, supporting the exogeneity of this shock in our context.

We count the cumulative number of local data scientist undergraduate programs established four years prior $(t-4)$ in the same state as the investor ($NumProgram$). We employ a two-stage least squares (2SLS) approach, using $NumProgram$ as the instrumental variable for the number of data scientists hired by the investor. This allows us to assess whether trading profitability of the investors increase with the positive shocks in data scientist hiring resulting from the higher supply of data scientist local talent pool. The findings of this analysis are presented in Table 4.

**[Insert Table 4 Here]**

Column (1) of Table 4 presents the result for the first-stage regression. The F-statistic exceeds 557, which suggests that we do not have a weak instrument problem. As expected, the instrumental variable $NumProgram$ is positively correlated with the number of data scientists hired by institutional investors located in the same states as the universities that established data science programs four or more years ago. The second-stage regression results are shown in column (2). The positive and significant coefficient for the instrumented number of data scientists indicates that trading profitability increases with the number of data scientists that are explained by an exogenous local data scientist supply. Columns

(3) to (8) confirm that this relationship holds consistently across all three types of data scientists. This evidence suggests that investors with more data scientists achieve higher trading profitability, indicating that data scientists help in gathering and analyzing valuable information for trading decisions, and this relationship is likely causal.

The next question is when investors gain the most advantage from the data scientists they hire. If all other investors holding the same stock also employ data scientists, the competition to extract useful information intensifies, reducing the advantage. In contrast, when fewer investors have data scientists, and the data scientist coverage is more concentrated, the investor with the most data scientists gains a more significant edge by accessing unique insights. Therefore, we would expect the advantage of having data scientists to be greater when data scientist coverage is concentrated, as opposed to being diluted across many investors.

We use an investor-firm-quarter level sample to examine the relationship between trading profitability and the concentration of data scientists covering the firm. The following regression is estimated:

$$\begin{aligned} Alpha_{i,j,t+1} = \beta_0 &+ \beta_1 PortfolioWeighted\ NumDS_{i,j,t} \times DS\ HHI_{j,t} + \beta_2 DS\ HHI_{j,t} \\ &+ \beta_3 PortfolioWeighted\ NumDS_{i,t} + Investor\ Controls + Firm\ Controls \\ &+ Fixed\ Effects, \end{aligned} \tag{5}$$

where $i$ denotes the 13F investor, $j$ denotes the stock, and $t$ denotes the quarter. Similar to Equation (3), the dependent variables represent the portfolio-weight change $\times$ alphas of the *stock* based on the CAPM, Fama-French 3-factor, and Fama-French 4-factor models. For example, CAPM $\alpha_{i,j,t+1}$ is calculated as:

$$CAPM\ \alpha_{i,j,t+1} = (Weight_{i,j,t} - Weight_{i,j,t-1}) \times CAPM\ \alpha_{j,t+1}, \tag{6}$$

where $Weight_{i,j,t}$ represents the portfolio weight the investor $i$ holds in stock $j$ in quater $t$ and $CAPM\alpha_{j,t+1}$ is the CAMP alpha of the stock $j$. The CAMP alpha of the stock $j$ is estimated at a monthly frequency, and we accumulate monthly alphas to compute quarterly alphas, which are then multiplied by 10,000 to be presented as basis points.

It is impractical to assume that investors allocate their data scientists' efforts equally across all the stocks in their portfolios. Stocks with higher portfolio weights naturally hold more importance, leading investors to allocate more resources toward gathering information on these stocks. To capture this, we construct an investor-stock-level data scientist measure, $PortfolioWeighted\ NumDS_{i,j,t}$, constructed by multiplying the portfolio weight of stock $j$ by the number of data scientists hired by the investor ($Weight_{i,j,t} \times NumDS_{i,t}$). $DS\ HHI$ represents the concentration of data scientists covering a firm, constructed in a similar manner to the Herfindahl index, as detailed in Section 2.1. Investor-level control variables are the same as those in Table 3. Firm-level control variables include the log of asset ($LogAsset$), Tobin's Q ($TobinQ$), ROA ($ROA$), and the log of firm age ($LogAge$).

In Equation (5), $\beta_3$ represents the effect of hiring data scientists on investor trading profitability when data scientist concentration is near zero, reflecting a situation where the distribution of data scientists covering the stock is highly dispersed. $\beta_2$ captures the effect of data scientist concentration on trading profitability for investors without any data scientists. The key coefficient, $\beta_1$, measures the impact of hiring data scientists on trading profitability, conditional on the level of data scientist concentration. A positive and significant $\beta_1$ suggests that the benefit of hiring data scientists on trading profitability is stronger when data scientist

concentration is higher.

**[Insert Table 5 Here]**

Table 5 presents the estimates for Equation (5). The coefficient on $DS\ HHI$ is negative and significant across all columns. This indicates that a higher concentration of data scientists is associated with lower trading profitability for investors without any data scientists. This aligns with the intuition that when only a few investors hold a "monopoly" on superior information through their data scientists, other investors are at an informational disadvantage, leading to poorer performance in trading. The key coefficients on the interaction term between $PortfolioWeighted\ NumDS$ and $DS\ HHI$ are positive and significant in most columns, suggesting that the positive impact of hiring data scientists on trading profitability is stronger when the concentration of data scientists is higher.

### 3.2 Strategic Investment and Hiring Decisions: How Do Financial Institutions Maintain Advantages Brought by Data Scientists?

Next, we explore how investors respond to the value that data scientists provide. First, we examine whether investors strategically tailor their portfolios to maximize the advantages provided by their data scientists.

Data scientists play a crucial role in uncovering valuable firm-specific information that helps investors make better trading decisions. If much of the data is unique to each firm—such as proprietary datasets, disclosures, and alternative data—it makes sense for investors with data scientists to concentrate their holdings on a smaller set of stocks. By doing so, data scientists can focus their efforts on a more limited number of companies, enabling a deeper anal-

ysis of each firm's individual dynamics and competitive positioning. This targeted approach allows investors to better leverage the unique data their data scientists analyze, leading to more informed investment decisions and maximizing the value that data scientists bring to the firm. In contrast, when data scientists generate firm-specific information, spreading data scientists' efforts across different firms too broadly would dilute their effectiveness, making it harder to extract firm-specific insights that drive trading profitability.

**[Insert Table 6 Here]**

To test whether investors maintain a concentrated portfolio when leveraging data scientists, we analyze the relationship between the portfolio holding Herfindahl index and the number of data scientists hired by the investor in column (1) of Table 6. The dependent variable in column (1) is $Holding\ HHI$, which measures the concentration of portfolio weights across the investor's holdings. The positive and significant coefficient on $NumDS$ suggests that investors who hire more data scientists tend to hold more concentrated portfolios. This indicates that investors allocate more of their portfolio to a smaller set of stocks, likely allowing data scientists to focus on extracting deeper insights from a targeted group of firms.

As shown in Table 5, the benefit of hiring data scientists on trading a given stock is amplified when the stock has a high data scientist concentration. Given this, it follows that investors with data scientists might choose to hold portfolios with stocks where data scientist concentration is higher. In column (2) of Table 6, we test this hypothesis by examining whether investors with data scientists tend to hold portfolios with stocks that exhibit higher data scientist concentration. The dependent variable is $Portfolio\ DS\ HHI$, which

represents the portfolio-level average data scientists concentration that is computed as the holding value-weighted average of the stock-level data scientists concentration ($DS\ HHI$). The coefficient on $NumDS$ is positive and significant in column (2), indicating that investors do indeed leverage their data scientists by strategically holding portfolios of stocks with higher data scientist concentration.[8] This allows them to benefit from the information "monopoly" effect highlighted in Table 5, where concentrated data scientist coverage among fewer investors enhances trading profitability for investors with data scientists.

In addition to portfolio strategies, investors also make strategic moves in the labor market. Since data scientists play a crucial role in making better investment decisions and enhancing trading profitability, it is natural to expect that investors have strong incentives to sustain this advantage by continually seeking out data science talent. These incentives grow even stronger when rivals are seen to lead in data scientist hiring. This prompts other investors to keep up in the talent race, particularly if they feel they are falling behind in securing this critical advantage. Anecdotal evidence supports this view, showing that competition for data science talent is intense in the finance industry.[9]

To test whether investors seek to gain an advantage and keep pace in the talent race, we examine whether investors hire more data scientists when there is a larger gap between the number of data scientists they employ and the number employed by their leading peer investor.

[8]This result is not mechanical. When institutions hire more data scientists, the average concentration of data scientists in their portfolios may go down. This depends on how institutions rebalance their portfolios and how data scientists are distributed across other institutions holding the same stocks.

[9]See, `https://darwinx-index.ams3.cdn.digitaloceanspaces.com/public-index/data_preview/BANKING/Report/DX_Data&AI_HumanCapitalReport_BankingEdition_v1.pdf`

$$\Delta NumDS_{i,t+2} = \beta_0 + \beta_1 NumDS\ Diff_{i,t} + Investor\ Controls + Fixed\ Effects, \quad (7)$$

where $i$ denotes the investor and $t$ denotes the year. $\Delta NumDS_{i,t+2}$ is the change in the number of data scientists employed by investor $i$ from $t+1$ to $t+2$. The key independent variable, $NumDS\ Diff_{i,t}$, the gap between the number of data scientists employed by the leading competitor (the one with the most data scientists among institutions holding the same stock) and the number employed by investor $i$:

$$DS\ Diff_{i,t} = \sum_j [(NumDS_{j,t} - NumDS_{i,t}) \times DS\ HHI_{j,t} \times Weight_{i,j,t}] \quad (8)$$

where $NumDS_{j,t}$ is the number of data scientists employed by the institution with the most data scientists holding stock $j$ (the leading competitor). $NumDS_{i,t}$ is the number of data scientists employed by the investor $i$. $Weight_{i,j,t}$ is a portfolio weight of stock $j$ in investor $i$'s portfolio.[10] The investor-level control variables are the same as those used in Table 3. Additionally, we include both investor and year fixed effects to account for time-invariant investor heterogeneity and the general time trend.

The coefficient $\beta_1$ in Equation (7) is expected to be positive and significant. When there is a larger gap between the investor $i$ and its competitors, the investor tends to keep up with the race and hire more data scientists in the future.

**[Insert Table 7 Here]**

[10]For the dependent variable in Equation (7), we examine the change in the number of data scientists from $t+1$ to $t+2$ (i.e., $\Delta NumDS_{i,t+2}$) instead of the change from $t$ to $t+1$ (i.e., $\Delta NumDS_{i,t+1} = NumDS_{i,t+1} - NumDS_{i,t}$). This is because if we examine the change from $t$ to $t+1$, the number of data scientists in $t$ ($NumDS_{i,t}$) can drive the dependent variable and our key independent variable in Equation (8) in the same direction, resulting in a mechanical relationship between the outcome variable and the independent variable.

Results in columns (1) and (2) of Table 7 align with our hypothesis. The positive and significant coefficients on $NumDS\ Diff$, both with and without control variables, indicate that investors with a larger gap in the number of data scientists relative to their competitors are more likely to recruit additional data scientists in the future to stay competitive in the talent race.

The effect is expected to be stronger when investors are more aware of their lag in data scientists relative to their competitors. If they hire data scientists who previously worked for their competitors, they are likely to gain better insights into their competitive disadvantage in terms of data scientist talent, leading to a stronger incentive to catch up in the talent race. In Column (3), we add the variable $Overlap\ NumDS\ Diff$, which captures the gap in the number of data scientists between the investor and its leading competitors, weighted by the number of data scientists hired by investor $i$ in year $t$ who previously worked for the leading competitor investor prior to year $t$:

$$Overlap\ NumDS\ Diff_{i,t} \tag{9}$$
$$= \sum_{j}[(NumDS_{j,t} - NumDS_{i,t}) \times DS\ HHI_{j,t} \times Weight_{i,j,t} \times Num\ Overlap\ DS_{i,j,t}],$$

where $Num\ Overlap\ DS_{i,j,t}$ denotes the number of data scientists hired by investor $i$ who previously worked for the leading competitor holding stock $j$. The result in Column (3) shows that investors are more aware of their data scientist gap and tend to hire more in the future when they employ data scientists who previously worked for their leading competitors.

One might be concerned that our results are simply driven by broader hiring trends in the industry, rather than reflecting a deliberate strategy to stay competitive in the tal-

ent race of data scientist. To address this, we create a placebo data scientist gap measure ($Placebo\ NumDs\ Diff$), replacing the number of data scientists hired by leading competitors with those hired by leading non-competitors who do not hold the same stock. Column (4) shows that this placebo measure is not significantly related to the focal investor's future data scientist hiring. This finding supports the idea that investors are specifically monitoring and responding to their competitors' information acquisition activities, rather than blindly following general industry hiring trends. These results highlight the strategic nature of investors' efforts to gain a competitive edge through data scientists.

### 3.3 Price Informativeness: What Does the Competition of Data Scientist Recruitment Bring to Capital Markets?

In Section 3.1, we show that investors with data scientists gain an information advantage, resulting in higher trading profitability, with the advantage being more pronounced when the data scientists covering the stocks are concentrated among a few investors. While this "information monopoly" benefits the trading profitability of these investors, it may come at the cost of overall market efficiency. When valuable information is concentrated within a few hands, the stock prices may not fully reflect all available information in the market. This is because a few investors with dominant information power lack the incentive to swiftly trade on their information (e.g., Kyle, 1985), and the broader investor base does not have access to the information possessed by those who have the information power. As a result, the information concentration could delay or distort the incorporation of new information into stock prices. This could affect the informational content of asset prices, resulting in lower price informativeness of the affected stocks. On the other hand, when multiple investors possess similar levels of information about a stock, they are likely to trade aggressively, leading to

prices quickly incorporating the fundamental information (e.g., Holden and Subrahmanyam, 1992).

Following Bai et al. (2016) and Kacperczyk et al. (2021), we measure price informativeness by examining how well current market prices predict future cash flows. Specifically, we estimate the following regression:

$$\begin{aligned} Earnings_{j,t+h} =& \beta_0 + \beta_1 LogMVA_{j,t} + \beta_2 LogMVA_{j,t} \times DS\ HHI_{j,t} + \beta_3 DS\ HHI_{j,t} \\ &+ \beta_4 LogMVA_{j,t} \times Log\ DS\ Coverage_{j,t} + \beta_5 Log\ DS\ Coverage_{j,t} \\ &+ LogMVA_{j,t} \times Firm\ Controls + Firm\ Controls + Fixed\ Effects, \end{aligned} \tag{10}$$

where $j$ denotes the firm and $t$ denotes the year. The dependent variable $Earnings_{j,t+h}$ represents the earnings of firm $j$ in year $t+h$, scaled by total assets in year $t$, with $h$ being one or three for short-term and long-term earnings. The independent variable $LogMVA_{j,t}$ is the natural logarithm of the ratio between the firm's market capitalization and total assets in year $t$, representing the stock prices in that year. $DS\ HHI_{j,t}$ is the data scientist concentration of firm $j$ in year $t$. $Log\ DS\ Coverage_{j,t}$ is the natural logarithm of the ownership-weighted average number of data scientists hired by investors of the stock. Firm-level control variables include institutional ownership as measured by the percentage of shares outstanding held by 13F institutional investors ($IO$), Herfindahl index of institutional holdings of the stock ($IO\ HHI$) (Kacperczyk et al., 2024; Xiong et al., 2024), total liability divided by total assets ($Leverage$), net property, plant, and equipment divided by total assets ($Tangibility$), cash divided by total assets ($Cash$), sales divided by total assets ($Sale$). We incorporate firm and year fixed effects in all test specifications to control for firm-specific characteristics that

remain constant over time and for overall time trends.

The coefficient $\beta_1$ measures the average level of price informativeness. If the stock market functions efficiently overall and stock prices incorporate relevant information that predicts future earnings, we expect $\beta_1$ to be positive and significant. The primary coefficient of interest is $\beta_2$, which captures the effect of data scientist concentration ($DS\ HHI_{j,t}$) on the sensitivity of future earnings to current stock prices. A positive and significant $\beta_2$ would indicate that a higher concentration of data scientists enhances price informativeness, while a negative and significant $\beta_2$ would suggest that such concentration impairs it.

In addition to examining the impact of data scientist concentration, we also investigate how data scientist coverage influences price informativeness. Intuitively, if more data scientists are covering a stock, while other factors remain constant, more information is likely to be extracted, leading to higher price informativeness. To account for the potential influence of other firm characteristics, we include interactions between $LogMVA_{j,t}$ and all control variables in our analysis.

**[Insert Table 8 Here]**

In column (1) of Panel A, Table 8, we examine the price informativeness for short-term earnings (one-year ahead earnings). As expected, the positive and significant coefficient on the interaction term between $LogMVA$ and $Log\ DS\ Coverage$ suggests that when more data scientists are covering a stock, the stock prices reflect more information relevant to future earnings, supporting the notion that data scientists indeed extract useful insights that enhance price formation.

However, the more critical finding is the negative and significant coefficient on the interaction term between $LogMVA$ and $DS\ HHI$. This result indicates that as data scientist concentration increases, the ability of stock prices to reflect information about future earnings declines. In other words, when data scientists are concentrated among a few investors, the information embedded in stock prices becomes less comprehensive, reducing its predictive power for future earnings. This highlights a critical efficiency cost of concentrated data scientist coverage, where valuable insights may not be fully integrated into the market, potentially leading to a less efficient price formation process. A one standard deviation increase in data scientist concentration is associated with a 10.97% decrease from the average price informativeness.[11]

The coefficients for the interaction terms between $LogMVA$ and the other control variables align with intuition. For instance, the positive and significant coefficient for the interaction between $LogMVA$ and $IO$ suggests that broader institutional investor coverage enhances information discovery. In contrast, the negative and significant coefficient for the interaction between $LogMVA$ and $IO\ HHI$ indicates that a more concentrated institutional investor base hampers information discovery.

The results in column (2) show similar effects for longer-term earnings. The negative and significant coefficient on the interaction term between $LogMVA$ and $DS\ HHI$ indicates that higher data scientist concentration negatively impacts price informativeness for predicting earnings three years ahead. This finding suggests that concentrated data scientist coverage not only affects the short-term predictability of stock prices but extends its adverse effects

[11] One standard deviation of $DS\ HHI$ $\times$ coefficient $\beta_2$ / price sensitivity of the average firm in our sample $= 0.205 \times (-0.023)/0.043 = -10.97\%$

to the anticipation of longer-term earnings as well.

We categorize data scientists into three groups based on their roles: data collection, data analytics, and data maintenance. In Panel B of Table 8, we analyze how the concentration of each group impacts price informativeness. The findings indicate that all three types of data scientists contribute to the observed effects. The coefficients for $\beta_2$ across all data scientist categories are consistently negative and significant, suggesting that higher concentration, regardless of the specific data scientist function, is consistently associated with lower price informativeness.

We demonstrate that data scientist concentration is negatively correlated with price informativeness. However, endogeneity remains a concern, as there could be factors influencing both the concentration of data scientists hired by institutional investors and the price informativeness of stocks.

To address this, we use mergers and acquisitions (M&As) among institutional investors as a source of exogenous variation in data scientist concentration. If M&A decisions are not motivated by the specific data scientist resources of acquirer and target institutions, these events can introduce exogenous variation in data scientist concentration of their portfolio firms. Furthermore, mergers lead to changes in data scientist concentration for portfolio firms, but these changes are not uniform in direction across all firms in the merged portfolio.

For instance, consider a hypothetical merger between two investors, Investor A and Investor B. Prior to the merger, Investor A employs 10 data scientists, while Investor B employs 5. Both investors hold stocks in a portfolio, which includes Stock 1 and Stock 2. Investor C, another large investor in Stock 1, employs significantly more data scientists. Pre-merger,

Investor C's dominant position in Stock 1 results in a high concentration of data scientists covering the stock, while the coverage for Stock 2, held only by Investors A and B, remains less concentrated. Post-merger, as Investors A and B combine their holdings and data scientists, the concentration of data scientists decreases for Stock 1, as the combined firm diversifies Investor C's dominance in coverage. However, for Stock 2, where the combined entity now holds a larger share and employs all the data scientists focused on this stock, the concentration increases. Therefore, depending on pre-merger conditions, the same merger can lead to divergent changes in data scientist concentration across different stocks within the portfolio. This heterogeneity makes it unlikely that M&As are motivated by the intent to alter the data scientist concentration of portfolio firms, providing a quasi-experimental setting to isolate and examine the causal impact of data scientist concentration on price informativeness.

We obtain M&A data from the Securities Data Company's (SDC) Mergers and Acquisitions Database, applying the following filters: (1) only completed transactions are considered, (2) transactions must be classified as "Merger," "Acquisition of Majority Interests," or "Acquisition of Assets," and (3) deal values must exceed $1 million. We then match the 13F institutional investors with the targets and acquirers listed in the SDC database using a name-matching algorithm, and manually verify each matched pair to ensure accuracy. These selection criteria yield a final sample of 20 deals over our sample period.

For the 20 deals, we define "affected stocks" as those held by either the acquirer or target at the time of the merger announcement. For each affected stock, we compute a hypothetical data scientist concentration by combining the fraction of shares held and the number of data

scientists employed by the acquirer and the target to form a combined entity. We then use the holdings the number of data scientists employed by and this combined entity and other investors in the same quarter to calculate a hypothetical data scientist concentration post-merger. By comparing this hypothetical concentration with the actual data scientist concentration for the stock, we determine whether the data scientist concentration decreases following the merger.

We define a dummy variable, *Down*, which takes the value of one for the post-merger period if the hypothetical data scientist concentration is lower than the actual concentration for the stock. Using a firm-year level sample, we retain data from three years before and three years after each merger for the affected stocks. We then examine whether a decrease in data scientist concentration resulting from mergers between financial institutions influences the price informativeness of these stocks.

**[Insert Table 9 Here]**

The results are presented in the columns (1) and (2) of Table 9. In both columns, the positive and significant coefficients on the interaction terms between $LogMVA$ and *Down* indicate that an exogenous decrease in data scientist concentration following financial institution mergers leads to improved price informativeness post-merger. This result holds regardless of whether data scientist coverage is included as a control variable. These findings provide causal evidence supporting our hypothesis that data scientist concentration negatively affects price informativeness.

In columns (3) and (4), we redefine the decrease in data scientist concentration ($Down$) based on the actual observed decrease in data scientist concentration post-merger, rather than the hypothetical decrease used in columns (1) and (2). The results remain consistent and even stronger, reinforcing our findings.

# 4 Conclusion

Our study sheds light on the critical role of data scientists in the financial industry, focusing on their impact on trading profitability, strategic decisions, and market efficiency. We document that institutional investors who employ more data scientists consistently achieve higher trading profitability, demonstrating that data scientists help institutions identify and capitalize on mispriced assets. By leveraging an exogenous increase in the local supply of data scientists, we establish a causal link between data scientist recruitment and improvements in trading performance, highlighting the value data scientists bring to portfolio management.

Beyond individual trading advantages, we find that investors make strategic adjustments in both portfolio allocation and talent acquisition to maximize the benefits provided by their data scientists. Investors tend to hold more concentrated portfolios, allowing data scientists to generate deeper insights into specific stocks. Additionally, we observe a competitive race in the labor market, where investors actively respond to their rivals' hiring activities, particularly when they perceive themselves as lagging in data scientist recruitment. These strategic moves underscore the significance of data scientists as a competitive asset in the capital market.

However, our findings also reveal a trade-off between the informational advantages gained

by individual investors and broader market efficiency. When data scientists are concentrated among a few institutional investors, stock prices become less informative about future earnings. This concentration creates an "information monopoly," where valuable insights are not fully reflected in stock prices, leading to inefficiencies in price formation. By exploiting exogenous variations in data scientist concentration resulting from mergers and acquisitions among institutional investors, we confirm that decreases in data scientist concentration improve price informativeness, further supporting our hypothesis.

Overall, our paper contributes to the growing literature on the role of big data and labor dynamics in financial markets. We emphasize the unique position of data scientists, who, unlike traditional data assets, cannot be shared across institutional investors simultaneously. This distributional concentration plays a critical role in shaping market outcomes and investor behavior. Our findings have implications for understanding how financial institutions harness human capital to gain a competitive edge and the potential consequences of these strategies for market efficiency.

## References


Bai, J., Philippon, T., Savov, A., 2016. Have financial markets become more informative? Journal of Financial Economics 122, 625–654.

Bonelli, M., Foucault, T., 2023. Displaced by big data: Evidence from active fund managers. Available at SSRN 4527672 .

Boustanifar, H., Grant, E., Reshef, A., 2017. Wages and Human Capital in Finance: International Evidence, 1970–2011*. Review of Finance 22, 699–745.

Cai, W., Chen, Y., Rajgopal, S., Azinovic-Yang, L., 2024. Diversity targets. Review of Accounting Studies 29, 2157–2208.

Chi, F., Hwang, B.-H., Zheng, Y., 2024. The use and usefulness of big data in finance: Evidence from financial analysts. Management Science 0, null.

Célérier, C., Vallée, B., 2019. Returns to Talent and the Finance Wage Premium. The Review of Financial Studies 32, 4005–4040.

Dessaint, O., FOUCAULT, T., FRESARD, L., 2024. Does alternative data improve financial forecasting? the horizon effect. The Journal of Finance 79, 2237–2287.

Dugast, J., Foucault, T., 2018. Data abundance and asset price informativeness. Journal of Financial Economics 130, 367–391.

Dugast, J., Foucault, T., 2024. Equilbruim data mining and data abundance. Journal of Finance, forthcoming .

Ellul, A., Pagano, M., Scognamiglio, A., 2019. Career Risk and Market Discipline in Asset Management. The Review of Financial Studies 33, 783–828.

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

Farboodi, M., Mihet, R., Philippon, T., Veldkamp, L., 2019. Big data and firm dynamics. AEA Papers and Proceedings 109, 38–42.

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

Holden, C. W., Subrahmanyam, A., 1992. Long-lived private information and imperfect competition. The Journal of Finance 47, 247–270.

Huang, S., Xiong, Y., Yang, L., 2022. Skill acquisition and data sales. Management Science 68, 6116–6144.

Jiang, H., Zheng, L., 2018. Active fundamental performance. The Review of Financial Studies 31, 4688–4719.

Kacperczyk, M., Nieuwerburgh, S. V., Veldkamp, L., 2014. Time-varying fund manager skill. The Journal of Finance 69, 1455–1484.

Kacperczyk, M., Nosal, J., Sundaresan, S., 2024. Market power and price informativeness. Review of Economic Studies p. rdae077.

Kacperczyk, M., Sundaresan, S., Wang, T., 2021. Do foreign institutional investors improve price efficiency? The Review of Financial Studies 34, 1317–1367.

Kumar, N., Mullally, K., Ray, S., Tang, Y., 2020. Prime (information) brokerage. Journal of Financial Economics 137, 371–391.

Kyle, A. S., 1985. Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society pp. 1315–1335.

Li, Q., Lourie, B., Nekrasov, A., Shevlin, T., 2022. Employee turnover and firm performance: Large-sample archival evidence. Management Science 68, 5667–5683.

Liang, C., Lourie, B., Nekrasov, A., Yoo, I. S., 2023. Voluntary disclosure of workforce gender diversity. Available at SSRN 3971818 .

Martin, I. W., Nagel, S., 2022. Market efficiency in the age of big data. Journal of Financial Economics 145, 154–177.

Oyer, P., 2008. The making of an investment banker: Stock market shocks, career choice, and lifetime income. The Journal of Finance 63, 2601–2628.

Philippon, T., Reshef, A., 2012. Wages and Human Capital in the U.S. Finance Industry: 1909–2006*. The Quarterly Journal of Economics 127, 1551–1609.

Veldkamp, L., 2023. Valuing data as an asset. Review of Finance 27, 1545–1562.

Xiong, Y., Yang, L., Zheng, Z., 2024. Institutional ownership concentration and informational efficiency. Available at SSRN .

Zhu, C., 2019. Big Data as a Governance Mechanism. The Review of Financial Studies 32, 2021–2061.

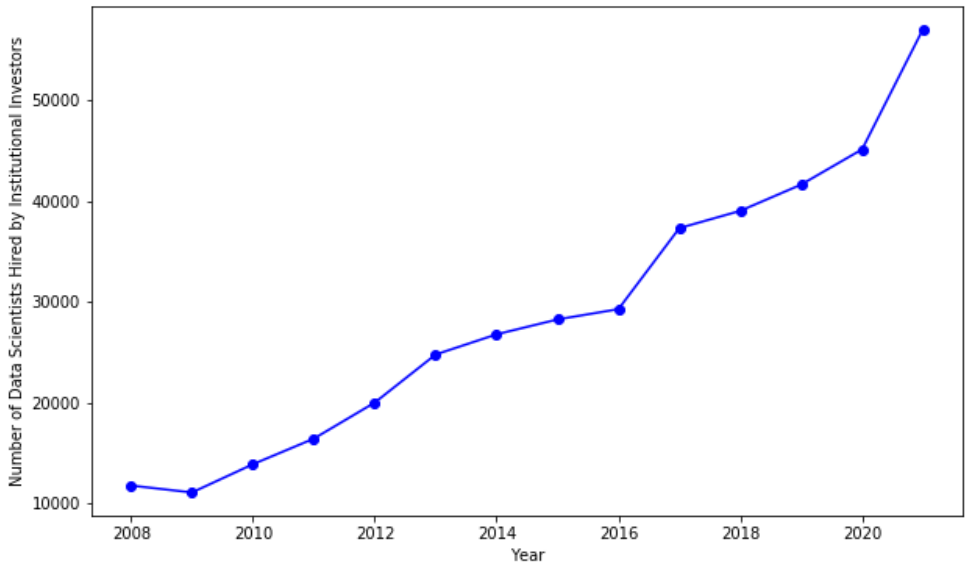


Figure 1. Number of Data Scientists Hired by Institutional Investors

This figure plots the time series for the total number of data scientists hired by all institutional investors on an annual frequency.

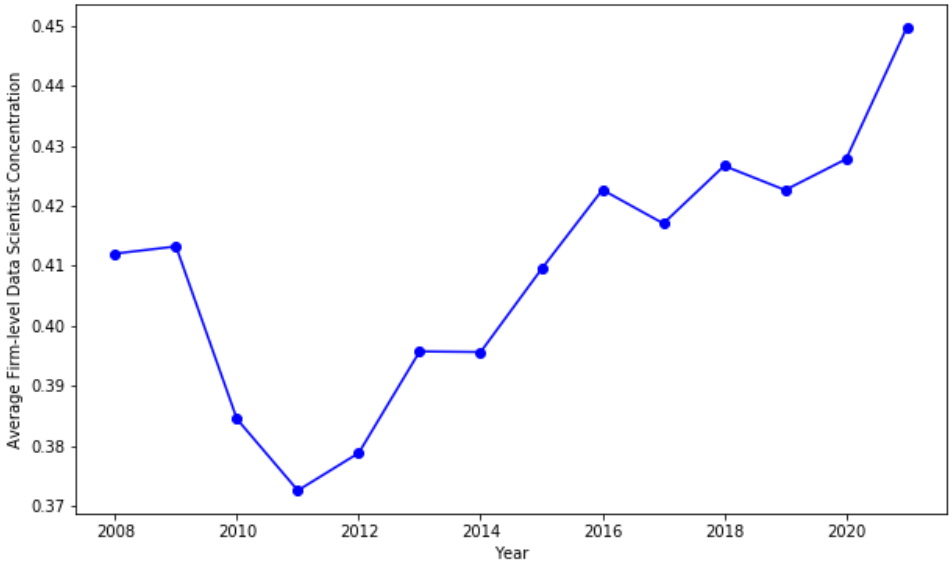


Figure 2. Average Firm-level Data Scientist Concentration

This figure plots the time-series patterns for the mean of data scientist concentration across all stocks, where the data scientist concentration is defined as Herfindahl index in Equation (2).

Table 1. Summary Statistics

This table presents summary statistics of our study. Panels A and B provide summary statistics for the institutional investor-level sample and firm-level sample, respectively. Our sample covers a period from 2008 to 2021. We exclude firms in finance (SIC 6000-6999) industries for firm-level samples.

Panel A. Institutional Investor-Level Summary Statistics

| VarName | Mean | SD | P25 | Median | P75 |
|---|---|---|---|---|---|
| NumDS | 2.052 | 10.107 | 0.000 | 0.000 | 0.000 |
| NumAnalysis | 1.182 | 5.884 | 0.000 | 0.000 | 0.000 |
| NumCollect | 0.233 | 1.135 | 0.000 | 0.000 | 0.000 |
| NumMaintain | 0.596 | 3.139 | 0.000 | 0.000 | 0.000 |
| NumProgram | 0.591 | 0.922 | 0.000 | 0.000 | 1.000 |
| CAPM $\alpha_{t+1}$ | 0.031 | 0.929 | -0.213 | 0.023 | 0.291 |
| FF3 $\alpha_{t+1}$ | 0.017 | 0.905 | -0.217 | 0.013 | 0.259 |
| FF4 $\alpha_{t+1}$ | 0.008 | 0.863 | -0.216 | 0.015 | 0.256 |
| LogTNA | 19.978 | 1.702 | 18.885 | 19.678 | 20.847 |
| Turnover | 0.126 | 0.189 | 0.028 | 0.067 | 0.152 |
| FewIndDummy | 0.022 | 0.146 | 0.000 | 0.000 | 0.000 |
| Log Market Cap | 26.602 | 13.792 | 19.990 | 26.775 | 33.300 |
| Volume | 0.099 | 0.073 | 0.058 | 0.088 | 0.123 |
| Gross Profit | -0.219 | 13.077 | 0.136 | 0.409 | 0.519 |

Panel B. Firm-Level Summary Statistics

| VarName | Mean | SD | P25 | Median | P75 |
|---|---|---|---|---|---|
| DS Coverage | 17.187 | 39.641 | 3.584 | 9.609 | 19.454 |
| DS Coverage Analysis | 11.218 | 26.413 | 2.224 | 6.140 | 12.802 |
| DS Coverage Collect | 1.398 | 2.853 | 0.399 | 0.959 | 1.625 |
| DS Coverage Maintain | 4.572 | 11.117 | 0.874 | 2.404 | 5.003 |
| DS HHI | 0.409 | 0.205 | 0.254 | 0.357 | 0.520 |
| Analysis HHI | 0.414 | 0.204 | 0.259 | 0.364 | 0.524 |
| Collect HHI | 0.402 | 0.225 | 0.232 | 0.337 | 0.521 |
| Maintain HHI | 0.436 | 0.207 | 0.280 | 0.385 | 0.550 |
| $\text{Earnings}_{t+1}$ | -0.018 | 0.271 | -0.039 | 0.054 | 0.111 |
| LogMVA | -0.156 | 1.256 | -0.764 | -0.036 | 0.637 |
| IO | 0.432 | 0.245 | 0.228 | 0.457 | 0.621 |
| IO HHI | 0.090 | 6.568 | 0.006 | 0.016 | 0.029 |
| LogAsset | 6.614 | 2.175 | 5.035 | 6.576 | 8.112 |
| Leverage | 0.511 | 0.283 | 0.302 | 0.501 | 0.676 |
| Tangibility | 0.504 | 0.453 | 0.146 | 0.358 | 0.784 |
| Cash | 0.177 | 0.205 | 0.036 | 0.104 | 0.231 |
| Sale | 0.839 | 0.707 | 0.344 | 0.675 | 1.138 |

Table 2. Institutional and Firm Characteristics, Data Scientist Hiring and Coverage

This table examines the relationship between institutional investor characteristics and their hiring of data scientists in Panel A, and the relationship between firm characteristics and data scientist coverage in Panel B. The dependent variable in Panel A, *Num DS*, represents the total number of data scientists employed by institutional investors in the year. The independent variables capture various institutional investor characteristics, including asset size (*LogTNA*), number of industries covered by the portfolio (*Log Number of SIC*), portfolio turnover (*Turnover*), number of local data scientist undergraduate programs (NumProgram), the types of investors (HedgeFund, PensionFund, Bank), and the total number of employees hired by the investor (NumEmployee). Columns (1), (2), and (3) progressively introduce fixed effects: column (1) includes no fixed effects, column (2) includes year fixed effects, and column (3) includes investor type by year fixed effects. Standard errors are clustered at the investment style-by-year level and reported in parentheses. In Panel B, we analyze the relationship between firm characteristics and the level of data scientist coverage and concentration. The dependent variable is *DS Coverage*, the ownership-weighted average number of data scientists covering a firm's stock. All independent variables are contemporaneously measured, with their definitions found in Appendix Table A1. Standard errors are clustered at the institution level in Panel A and the stock level in Panel B. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

Panel A. Institutional Investor Characteristics and the Hiring of Data Scientists

| | (1) | (2) | (3) |
|---|---|---|---|
| | | Num DS | |
| LogTNA | -0.038 | -0.075 | -0.071 |
| | (0.122) | (0.127) | (0.139) |
| Num NonDS Employee | 0.041*** | 0.041*** | 0.042*** |
| | (0.000) | (0.001) | (0.001) |
| Log Number of SIC | 0.796*** | 0.852*** | 0.744** |
| | (0.172) | (0.254) | (0.313) |
| Turnover | 2.771*** | 2.927** | 3.672*** |
| | (0.979) | (1.264) | (1.293) |
| HedgeFund | 0.859** | 0.970*** | |
| | (0.397) | (0.311) | |
| PensionFund | -18.391*** | -18.293*** | |
| | (1.821) | (4.883) | |
| Bank | -3.099*** | -2.961*** | |
| | (0.935) | (0.711) | |
| NumProgram | 0.732*** | 0.599*** | 0.618*** |
| | (0.194) | (0.199) | (0.200) |
| Institution Type × Time FEs | × | × | ✓ |
| Time FEs | × | ✓ | × |
| Obs. | 49,617 | 49,617 | 49,601 |
| Adj. R2 | 0.885 | 0.885 | 0.888 |

Panel B. Firm Characteristics and Data Scientist Coverage

| | (1)<br>DS Coverage | (2)<br>DS Coverage |
|---|---|---|
| IO | 35.956*** | 38.808*** |
| | (3.842) | (4.315) |
| IO HHI | 6.168 | 5.220 |
| | (4.972) | (4.460) |
| LogAsset | 0.873*** | 0.413 |
| | (0.210) | (0.263) |
| TobinQ | -0.425*** | -0.656*** |
| | (0.160) | (0.212) |
| LogAge | -1.615*** | -1.250*** |
| | (0.464) | (0.482) |
| Leverage | 2.882** | 3.772*** |
| | (1.232) | (1.413) |
| Tangibility | 1.743* | 2.435* |
| | (0.908) | (1.351) |
| Cash | 6.436*** | 6.424*** |
| | (1.688) | (1.959) |
| Sale | -0.119 | -1.452* |
| | (0.401) | (0.854) |
| HasFraud | 3.764** | 2.894* |
| | (1.570) | (1.552) |
| SIC×Time FEs | × | ✓ |
| Time FEs | ✓ | × |
| Obs. | 46,681 | 45,873 |
| Adj. R2 | 0.091 | 0.071 |

Table 3. Institutional Investor' Trading Profitability Performance and Data Scientist Advantage

This table presents the baseline results and a detailed breakdown of the effect of data scientist hiring on investors' trading profitability in Panel A and B, respectively. In Panel A, we examine the overall impact of data scientist hiring on trading profitability. Panel B extends the analysis of the relationship between data scientist hiring and trading profitability by dividing data scientists into three categories: data scientists for data analysis, data scientists for data collection, and data scientists for data maintenance. The dependent variables include CAPM $\alpha_{t+1}$, FF3 $\alpha_{t+1}$, and FF4 $\alpha_{t+1}$, representing the position change-weighted alphas of the portfolio based on the CAPM, Fama-French 3-factor, and Fama-French 4-factor models, respectively. CAPM $\alpha_{t+1}$, for instance, is calculated as $CAPM\ \alpha_{t+1} = \sum_j (Weight_{i,j,t} - Weight_{i,j,t-1}) \times CAPM\ \alpha_{j,t+1}$, where $Weight_{i,j,t}$ represents the portfolio weight the investor $i$ holds in stock $j$ in quater $t$ and $CAPM\alpha_{j,t+1}$ is the CAMP alpha of the stock $j$. Investor control variables, with definitions provided in the Appendix Table A1, are included in all test specifications. The sample is at the investor-year-quarter level, and both investor and year-quarter fixed effects are applied. Standard errors are clustered at the investor level and reported in parentheses. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

Panel A. The Overall Impact of Data Scientist Hiring on Trading Profitability

| | (1) CAPM $\alpha_{t+1}$ | (2) FF3 $\alpha_{t+1}$ | (3) FF4 $\alpha_{t+1}$ |
|---|---|---|---|
| NumDS | 0.004*** | 0.002** | 0.002*** |
| | (0.001) | (0.001) | (0.001) |
| LogTNA | 0.034*** | 0.033*** | 0.036*** |
| | (0.005) | (0.004) | (0.004) |
| Log Number of Firm | 0.002 | -0.004 | -0.011* |
| | (0.006) | (0.005) | (0.005) |
| Turnover | -0.026 | -0.002 | -0.064** |
| | (0.028) | (0.028) | (0.027) |
| FewIndDummy | -0.009 | 0.005 | 0.016 |
| | (0.020) | (0.019) | (0.019) |
| Log Market Cap | -0.000 | -0.001 | -0.000 |
| | (0.000) | (0.000) | (0.000) |
| Volume | -0.238*** | 0.024 | -0.142** |
| | (0.060) | (0.056) | (0.057) |
| Gross Profit | 0.000 | -0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| Investor FEs | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ |
| Obs. | 186,478 | 186,478 | 186,478 |
| Adj. R2 | 0.087 | 0.091 | 0.090 |

Panel B. Different Types of Data Scientist on Investors' Trading Profitability

| | (1) CAPM $\alpha_{t+1}$ | (2) CAPM $\alpha_{t+1}$ | (3) CAPM $\alpha_{t+1}$ | (4) CAPM $\alpha_{t+1}$ |
|---|---|---|---|---|
| NumDS Analysis | 0.006*** | | | 0.005*** |
| | (0.001) | | | (0.001) |
| NumDS Collect | | 0.012** | | -0.003 |
| | | (0.006) | | (0.007) |
| NumDS Maintain | | | 0.009*** | 0.004* |
| | | | (0.002) | (0.002) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Investor FEs | ✓ | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ | ✓ |
| Obs. | 186,478 | 186,478 | 186,478 | 186,478 |
| Adj. R2 | 0.087 | 0.087 | 0.087 | 0.087 |

Table 4. Local Data Scientist Undergraduate Programs, Data Scientists Hiring, and Investor Performance

This table presents the results of a two-stage least squares (2SLS) regression that examines the impact of exogenous increases in local data scientist undergraduate programs on investor hiring of data scientists and subsequent trading profitability. The instrument, NumProgram, represents the cumulative number of local data scientist undergraduate programs established four years prior $(t-4)$ in the same state as the investor. The dependent variable in the second stage is CAPM $\alpha_{t+1}$, which measures the trading profitability of the investor. Columns (1) and (2) focus on the total number of data scientists (NumDS), with column (1) showing the first-stage results where NumProgram instruments for NumDS, and column (2) showing the second-stage results. Columns (3) through (8) adopt a similar framework, with NumProgram instrumenting for different categories of data scientists. $\widehat{\text{Variable}}$ represents a variable instrumented by the instrument variable in the first-stage regressions. Investor control variables, as defined in the Appendix, are included. Investor and year fixed effects are included. Standard errors are clustered at the investor level and reported in parentheses. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

| | (1) First Stage | (2) Second Stage | (3) First Stage | (4) Second Stage | (5) First Stage | (6) Second Stage | (7) First Stage | (8) Second Stage |
|---|---|---|---|---|---|---|---|---|
| | NumDS | CAPM $\alpha_{t+1}$ | NumDS Analysis | CAPM $\alpha_{t+1}$ | NumDS Collect | CAPM $\alpha_{t+1}$ | NumDS Maintain | CAPM $\alpha_{t+1}$ |
| $\widehat{\text{NumDS}}$ | | 0.095*** | | | | | | |
| | | (0.019) | | | | | | |
| $\widehat{\text{NumDS Analysis}}$ | | | | 0.182*** | | | | |
| | | | | (0.040) | | | | |
| $\widehat{\text{NumDS Collect}}$ | | | | | | 2.374** | | |
| | | | | | | (1.044) | | |
| $\widehat{\text{NumDS Maintain}}$ | | | | | | | | 0.233*** |
| | | | | | | | | (0.046) |
| NumProgram | 0.327*** | | 0.170*** | | 0.013** | | 0.133*** | |
| | (0.052) | | (0.031) | | (0.006) | | (0.021) | |
| LogTNA | 0.319*** | 0.002 | 0.191*** | -0.003 | 0.015** | -0.005 | 0.113*** | 0.006 |
| | (0.054) | (0.009) | (0.030) | (0.010) | (0.006) | (0.023) | (0.020) | (0.008) |
| Log Number of Firm | 0.098 | -0.006 | 0.055 | -0.007 | 0.014 | -0.030 | 0.022 | -0.002 |
| | (0.076) | (0.009) | (0.046) | (0.010) | (0.009) | (0.026) | (0.026) | (0.008) |
| Turnover | 0.147 | -0.039 | 0.078 | -0.039 | -0.004 | -0.015 | 0.078* | -0.043 |
| | (0.100) | (0.030) | (0.055) | (0.030) | (0.012) | (0.039) | (0.040) | (0.030) |
| FewIndDummy | 0.577*** | -0.065** | 0.263*** | -0.059** | 0.052** | -0.133* | 0.229*** | -0.064** |
| | (0.148) | (0.027) | (0.101) | (0.029) | (0.021) | (0.075) | (0.052) | (0.025) |
| Log Market Cap | 0.003 | -0.000 | 0.002 | -0.000 | 0.000 | -0.001 | 0.000 | -0.000 |
| | (0.005) | (0.001) | (0.003) | (0.001) | (0.001) | (0.002) | (0.002) | (0.001) |
| Volume | -1.830*** | -0.041 | -0.927*** | -0.045 | -0.122*** | 0.075 | -0.740*** | -0.041 |
| | (0.401) | (0.076) | (0.209) | (0.077) | (0.045) | (0.160) | (0.197) | (0.082) |
| Gross Profit | -0.000** | 0.000 | -0.000** | 0.000 | -0.000* | 0.000 | -0.000* | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Investor FEs | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Obs. | 186,478 | 186,478 | 186,478 | 186,478 | 186,478 | 186,478 | 186,478 | 186,478 |
| F-statistic | 557.998 | | 419.085 | | 60.290 | | 648.098 | |

Table 5. Institutional Investor' Trading Profitability and Firm Data Scientist Concentration

This table examines the relationship between an investor's data scientist advantage in trading profitability and the concentration of investors' data scientists at the firms in which they invest. The dependent variables include CAPM $\alpha_{i,j,t+1}$, FF3 $\alpha_{i,j,t+1}$, and FF4 $\alpha_{i,j,t+1}$, representing the position change-weighted alphas of the investment in stock $j$ based on the CAPM, Fama-French 3-factor, and Fama-French 4-factor models, respectively. For example, CAPM $\alpha_{i,j,t+1}$ is calculated as: $CAPM\ \alpha_{i,j,t+1} = (Weight_{i,j,t} - Weight_{i,j,t-1}) \times CAPM\alpha_{j,t+1}$, where $Weight_{i,j,t}$ represents the portfolio weight the investor $i$ holds in stock $j$ in quater $t$ and $CAPM\ \alpha_{j,t+1}$ is the CAMP alpha of the stock $j$. Portfolio-weighted NumDS is the number of data scientists hired by investor $i$, weighted by the portfolio weight of stock $j$ in investor $i$'s portfolio. DS HHI represents the concentration of data scientists for stock $j$. Investor and firm control variables, with definitions provided in the Appendix Table A1, are included in all test specifications. The sample is at the investor-firm-year-quarter level, and both investor, firm, and year-quarter fixed effects are applied. Standard errors are clustered at the investor level and reported in parentheses. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

| | (1) CAPM $\alpha_{i,j,t+1}$ | (2) FF3 $\alpha_{i,j,t+1}$ | (3) FF4 $\alpha_{t+1}$ |
|---|---|---|---|
| Portfolio-Weighted NumDS × DS HHI | 0.017 | 0.017** | 0.018*** |
| | (0.013) | (0.007) | (0.006) |
| DS HHI | -0.103*** | -0.104*** | -0.114*** |
| | (0.010) | (0.010) | (0.010) |
| Portfolio-Weighted NumDS | -0.009 | -0.006 | -0.008 |
| | (0.012) | (0.007) | (0.006) |
| LogTNA | -0.004 | 0.005 | 0.001 |
| | (0.004) | (0.004) | (0.004) |
| Log Number of Firm | -0.007 | -0.007 | 0.005 |
| | (0.006) | (0.005) | (0.005) |
| Turnover | -0.100** | -0.025 | -0.098** |
| | (0.042) | (0.040) | (0.042) |
| FewIndDummy | -1.722** | -0.616 | -0.275 |
| | (0.751) | (0.704) | (0.667) |
| LogAsset | 0.141*** | 0.118*** | 0.121*** |
| | (0.007) | (0.006) | (0.006) |
| Tobin Q | 0.000*** | 0.000*** | 0.000*** |
| | (0.000) | (0.000) | (0.000) |
| ROA | -0.005*** | -0.004*** | -0.005*** |
| | (0.001) | (0.001) | (0.001) |
| LogAge | 0.014* | -0.041*** | -0.019** |
| | (0.008) | (0.008) | (0.008) |
| Investor FEs | ✓ | ✓ | ✓ |
| Firm FEs | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ |
| Obs. | 30,640,198 | 30,638,578 | 30,636,392 |
| Adj. R2 | 0.008 | 0.007 | 0.008 |

Table 6. Data Scientist Advantage and Portfolio-Level Data Scientist Concentration

This table examines how institutional investors leverage their data scientist advantage through portfolio choices. In column (1), the dependent variable is Holding HHI, which measures the Herfindahl-Hirschman Index (HHI) of portfolio weights across stocks held by the investor in quarter $t+1$. The dependent variable in column (2) is Portfolio DS HHI, which is the portfolio-weighted average of the Data Scientist Herfindahl-Hirschman Index (DS HHI) for the stocks held in the portfolio in quarter $t+1$. The DS HHI measures the concentration of data scientists across the institutions holding each stock, with higher values indicating more concentrated coverage. The key independent variable is NumDS, which is the number of data scientists hired by the investor. All test specifications are consistent with those in Table 3. Standard errors are clustered at the investor level and reported in parentheses. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

| | (1) Holding $HHI_{t+1}$ | (2) Portfolio DS $HHI_{t+1}$ |
|---|---|---|
| NumDS | 0.004*** | 0.004*** |
| | (0.001) | (0.001) |
| LogTNA | -0.690*** | 0.627*** |
| | (0.108) | (0.122) |
| Log Number of Firm | -4.861*** | -0.588*** |
| | (0.285) | (0.092) |
| Turnover | 0.839 | -1.217*** |
| | (0.577) | (0.368) |
| FewIndDummy | 15.607*** | 1.513** |
| | (0.908) | (0.584) |
| Log Market Cap | -0.019* | -0.090*** |
| | (0.010) | (0.015) |
| Volume | 0.496 | 0.457 |
| | (1.476) | (3.472) |
| Gross Profit | -0.005** | 0.001 |
| | (0.003) | (0.002) |
| Investor FEs | ✓ | ✓ |
| Time FEs | ✓ | ✓ |
| Obs. | 186,478 | 186,478 |
| Adj. R2 | 0.840 | 0.621 |

Table 7. Investor's Reaction to Competitor Data Scientist Hiring and Adjustments in Data Scientist Employment

This table explores how institutional investors adjust their data scientist hiring in response to the hiring activities of their competitors. The dependent variable is $\Delta NumDS_{i,t+2}$, representing the change in the number of data scientists employed by investor $i$ from $t+1$ to $t+2$. The key independent variable, $NumDS\ Diff_{i,t}$, captures the gap between the number of data scientists employed by the leading competitor (the institution with the most data scientists holding the same stock) and the number employed by investor $i$. The detailed definition is provided in Equation 8. In column (3), the variable *Overlap NumDS Diff* is added to account for the number of data scientists hired by investor $i$ in quarter $t$ who previously worked for the leading competitor before quarter $t$. Column (4) introduces a placebo variable, *Placebo NumDS Diff*, which measures the data scientist gap using non-competitor institutions (those that do not hold the same stock) instead of competitors. Investor and year fixed effects are included in all test specifications. Investor-level control variables are included, and standard errors are clustered at the investor style-year level. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | $\Delta$ NumDS$_{t+2}$ | | |
| NumDS Diff | 0.009** | 0.010** | 0.010** | 0.010** |
| | (0.004) | (0.004) | (0.004) | (0.004) |
| Overlap NumDS Diff | | | 0.075** | |
| | | | (0.037) | |
| Placebo NumDS Diff | | | | 0.001 |
| | | | | (0.010) |
| LogTNA | | -0.001 | -0.001 | -0.001 |
| | | (0.004) | (0.004) | (0.004) |
| Log Number of Firm | | 0.028** | 0.028** | 0.028** |
| | | (0.013) | (0.013) | (0.013) |
| Turnover | | 0.022 | 0.021 | 0.022 |
| | | (0.027) | (0.027) | (0.027) |
| FewIndDummy | | 0.034 | 0.036 | 0.035 |
| | | (0.052) | (0.052) | (0.052) |
| Log Market Cap | | 0.003** | 0.003** | 0.003** |
| | | (0.001) | (0.001) | (0.001) |
| Volume | | -0.190 | -0.184 | -0.189 |
| | | (0.129) | (0.129) | (0.129) |
| Gross Profit | | 0.000 | 0.000 | 0.000 |
| | | (0.000) | (0.000) | (0.000) |
| Investor FEs | ✓ | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ | ✓ |
| Obs. | 44,196 | 44,196 | 44,196 | 44,196 |
| Adj. R2 | 0.4099 | 0.4101 | 0.4110 | 0.4101 |

Table 8. Data Scientist Coverage, Concentration, and Stock Price Informativenesss

This table examines the relationship between the natural logrithm of one plus the number of data scientists covering a stock (*Log DS Coverage*), the concentration of data scientists (*DS HHI*), and the stock's price informativeness. Panel B extends the analysis of the relationship between data scientist coverage, concentration, and stock price informativeness by categorizing data scientists into three types: data scientists for data analysis, data scientists for data collection, and data scientists for data maintenance. The dependent variables are future earnings at $t+1$ and $t+3$ divided by total assets in year $t$ ($\text{Earnings}_{t+1}$ and $\text{Earnings}_{t+3}$). The independent variables include the natural logarithm of market capitalization divided by total assets (*LogMVA*) and its interaction with *Log DS Coverage* and *DS HHI*. In Panel B, *Log DS Coverage Analysis*, *Log DS Coverage Collect*, and *Log DS Coverage Maintain* are the natural logrithm of number of data scientists for data analysis, data scientists for data collection, and data scientists for data maintenance that cover the firm. Firm-level control variables, as well as the interactions between these controls and LogMVA, are included in the regressions. Firm and year fixed effects are applied, and standard errors are clustered at the firm level and reported in parentheses. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

Panel A. Data Scientist Coverage, Concentration, and Stock Price Informativeness

| | (1) $\text{Earnings}_{t+1}$ | (2) $\text{Earnings}_{t+3}$ |
|---|---|---|
| LogMVA | 0.039*** | 0.032*** |
| | (0.006) | (0.012) |
| LogMVA × DS HHI | -0.023*** | -0.043*** |
| | (0.006) | (0.011) |
| DS HHI | 0.020** | -0.021 |
| | (0.008) | (0.016) |
| LogMVA × Log DS Coverage | 0.006*** | 0.011*** |
| | (0.001) | (0.002) |
| LogMVA × IO | 0.027*** | 0.039*** |
| | (0.006) | (0.010) |
| LogMVA × IO HHI | -0.042*** | -0.043*** |
| | (0.008) | (0.016) |
| LogMVA × Leverage | -0.060*** | -0.051*** |
| | (0.005) | (0.010) |
| LogMVA × Tangibility | 0.015*** | -0.044*** |
| | (0.004) | (0.006) |
| LogMVA × Cash | -0.102*** | -0.131*** |
| | (0.009) | (0.019) |
| LogMVA × Sale | 0.028*** | 0.022*** |
| | (0.002) | (0.004) |
| Log DS Coverage | 0.006*** | 0.019*** |
| | (0.002) | (0.004) |
| IO | 0.003 | 0.020 |
| | (0.011) | (0.021) |
| IO HHI | -0.044*** | -0.061** |
| | (0.014) | (0.029) |
| LogAsset | 0.067*** | 0.082*** |
| | (0.004) | (0.009) |
| Leverage | -0.020** | -0.066*** |
| | (0.009) | (0.019) |
| Tangibility | -0.027*** | -0.038* |
| | (0.011) | (0.020) |
| Cash | 0.057*** | 0.067** |
| | (0.013) | (0.028) |
| Sale | 0.096*** | 0.103*** |
| | (0.006) | (0.011) |
| Firm FEs | ✓ | ✓ |
| Time FEs | ✓ | ✓ |
| Obs. | 46,146 | 34,947 |
| Adj. R2 | 0.8007 | 0.7303 |

Panel B. Data Scientist Coverage, Concentration, and Stock Price Informativeness: Different Types of Data Scientist

| | (1) Earnings$_{t+1}$ | (2) Earnings$_{t+3}$ | (3) Earnings$_{t+1}$ | (4) Earnings$_{t+3}$ | (5) Earnings$_{t+1}$ | (6) Earnings$_{t+3}$ |
|---|---|---|---|---|---|---|
| LogMVA | 0.041*** | 0.034*** | 0.046*** | 0.044*** | 0.044*** | 0.039*** |
| | (0.006) | (0.011) | (0.006) | (0.011) | (0.006) | (0.011) |
| LogMVA × Analysis HHI | -0.022*** | -0.039*** | | | | |
| | (0.006) | (0.012) | | | | |
| Analysis HHI | 0.025*** | -0.016 | | | | |
| | (0.008) | (0.016) | | | | |
| LogMVA × Log DS Coverage Analysis | 0.006*** | 0.010*** | | | | |
| | (0.001) | (0.003) | | | | |
| Log DS Coverage Analysis | 0.005** | 0.018*** | | | | |
| | (0.002) | (0.004) | | | | |
| LogMVA × Collect HHI | | | -0.022*** | -0.040*** | | |
| | | | (0.006) | (0.012) | | |
| Collect HHI | | | 0.009 | -0.023 | | |
| | | | (0.008) | (0.016) | | |
| LogMVA × Log DS Coverage Collect | | | 0.008*** | 0.014*** | | |
| | | | (0.002) | (0.004) | | |
| Log DS Coverage Collect | | | 0.010** | 0.026*** | | |
| | | | (0.004) | (0.007) | | |
| LogMVA × Maintain HHI | | | | | -0.021*** | -0.032*** |
| | | | | | (0.006) | (0.011) |
| Maintain HHI | | | | | 0.013* | -0.016 |
| | | | | | (0.008) | (0.015) |
| LogMVA × Log DS Coverage Maintain | | | | | 0.005*** | 0.009*** |
| | | | | | (0.001) | (0.003) |
| Log DS Coverage Maintain | | | | | 0.006** | 0.019*** |
| | | | | | (0.003) | (0.005) |
| Controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Firm FEs | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Obs. | 46,146 | 34,947 | 46,146 | 34,947 | 46,146 | 34,947 |
| Adj. R2 | 0.8007 | 0.7301 | 0.8004 | 0.7299 | 0.8005 | 0.7299 |

Table 9. Exogenous Decrease in Data Scientist Concentration After Institutional Investor Mergers and Price Informativeness

This table explores the impact of an exogenous decrease in data scientist Herfindahl-Hirschman Index (DS HHI) following mergers and acquisitions between institutional investors on stock price informativeness. The sample includes observations from three years before and three years after the merger. The key independent variable, *Down*, is a dummy variable equal to one for the stock that experiences a decrease in DS HHI post-merger. In columns (1) and (2), *Down* represents a hypothetical decrease in DS HHI based on pre-merger data, while in columns (3) and (4), *Down* captures the actual decrease in DS HHI post-merger. Firm and year fixed effects are included, and standard errors are clustered at the firm level. ***, **, and * indicate the 1%, 5%, and 10% levels of statistical significance, respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Earnings$_{t+1}$ | | | |
| | Hypothetical DS HHI Down | | Actual DS HHI Down | |
| LogMVA | 0.049*** | 0.048*** | 0.049*** | 0.047*** |
| | (0.006) | (0.007) | (0.006) | (0.007) |
| LogMVA × Down | 0.001* | 0.002* | 0.002** | 0.002** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Down | -0.001 | -0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| LogMVA × Log DS Coverage | | 0.001 | | 0.001 |
| | | (0.001) | | (0.001) |
| Log DS Coverage | | 0.002 | | 0.002 |
| | | (0.002) | | (0.002) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Firm FEs | ✓ | ✓ | ✓ | ✓ |
| Time FEs | ✓ | ✓ | ✓ | ✓ |
| Obs. | 37,986 | 37,986 | 37,986 | 37,986 |
| Adj. R2 | 0.8171 | 0.8171 | 0.8171 | 0.8171 |

# 5 Appendix

Table A1. Variable Definition

| Variable | Definition |
|---|---|
| Dependent variables | |
| $CAPM\ \alpha_{t+1}$ | One-factor (Market model) alpha in the next quarter. |
| $FF3\ \alpha_{t+1}$ | Three-factor (Market, SMB, HML model) alpha in the next quarter. |
| $FF4\ \alpha_{t+1}$ | Four-factor (Market, SMB, HML, MOM model) alpha in the next quarter. |
| $Holding\ HHI$ | The Herfindahl index of investor's portfolio holdings in the end of year $t$. |
| $Prtf\ DS\ HHI$ | The weighted average $DS\ HHI$ of the investor's portfolio in the end of year $t$, weighted by the portfolio weight of each stock in the portfolio. |
| $\Delta NumDS_{t+2}$ | The difference between the number of data scientists hired by the investor in year $t+2$ and year $t+1$. |
| $Earnings_{t+1}$ | EBIT in year $t+1$ divided by total assets in year $t$. |
| Independent variables | |
| $NumDS$ | The number of data scientists hired by the investor |
| *Log DS Coverage* | The natural logarithm of the ownership-weighted average number of data scientists hired by investors of the stock. |
| $DS\ HHI$ | Herfindahl index of ownership-weighted number of data scientists hired by investors of the stock. |
| *Log DS Coverage Analysis* | The natural logarithm of the ownership-weighted average number of data scientists specializing in data analytics hired by investors of the stock. |
| $Analysis\ HHI$ | Herfindahl index of ownership-weighted number of data scientists specializing in data analytics hired by investors of the stock. |
| *Log DS Coverage Collect* | The natural logarithm of the ownership-weighted average number of data scientists specializing in data collections hired by investors of the stock. |
| $Collect\ HHI$ | Herfindahl index of ownership-weighted number of data scientists specializing in data collection hired by investors of the stock. |
| *Log DS Coverage Maintain* | The natural logarithm of the ownership-weighted average number of data scientists specializing in data maintenance hired by investors of the stock. |
| $Maintain\ HHI$ | Herfindahl index of ownership-weighted number of data scientists specializing in data maintenance hired by investors of the stock. |
| $LogMVA$ | The natural logarithm of market capitalization divided by total assets. |

| Variable | Definition |
|---|---|
| $IO$ | Institutional ownership as measured by the percentage of shares outstanding held by 13F institutional investors. |
| $IO\ HHI$ | Herfindahl index of institutional holdings of the stock |
| $Leverage$ | Total liability divided by total assets. |
| $Tangibility$ | Net property, plant, and equipment divided by total assets |
| $Cash$ | Cash divided by total assets |
| $Sale$ | Sales divided by total assets |
| $LogAsset$ | The natural logarithm of total assets of the firm. |
| $LogTNA$ | The natural logarithm of total assets under management of the investor. |
| $Log\ Number\ of\ Firm$ | The natural logarithm of the number of stocks in the investor's portfolio. |
| $Turnover$ | Investor portfolio turnover during a quarter is calculated as the minimum of the absolute values of buys and sells made by investor $i$ during quarter $t$, divided by the average of the total portfolio value at the end of quarter $t-1$ and quarter $t$. |
| $FewIndDummy$ | Dummy variable that equals one if the investor's portfolio covers two or fewer industries. |
| $Log\ Market\ Cap$ | The ownership-weighted average of the natural logarithm of the market capitalization of stocks in the portfolio. |
| $Volume$ | The ownership-weighted average of the trading volumes of stocks in the portfolio. |
| $Gross\ Profit$ | The ownership-weighted average of the gross profit margin of stocks in the portfolio. |
| $NumDS\ Diff$ | $NumDS\ Diff_{i,t} = \sum_j [(NumDS_{j,t} - NumDS_{i,t}) \times DS\ HHI_{j,t} \times w_{i,j,t}]$ where $NumDS_{j,t}$ is the number of data scientists employed by the leading competitor holding stock $j$. $DS\ HHI_{j,t}$ is the HHI of data scientists for stock $j$. $w_{i,j,t}$ is a portfolio weight of stock $j$ in investor $i$'s portfolio. |
| $Overlap\ NumDS\ Diff$ | $Overlap\ DS\ Diff_{i,t} = \sum_j [(NumDS_{j,t} - NumDS_{i,t}) \times DS\ HHI_{j,t} \times w_{i,j,t} \times Num\ Overlap\ DS_{j,i,t}]$ where $Num\ Overlap\ DS_{j,i,t}$ denotes the number of data scientists hired by investor $i$ who previously worked for the leading competitor holding stock $j$. |
| $Placebo\ NumDS\ Diff$ | $DS\ Diff_{i,t} = \sum_j [(NumDS_{j,t} - NumDS_{i,t}) \times DS\ HHI_{j,t} \times w_{i,j,t}]$ where $NumDS_{j,t}$ is the number of data scientists employed by a random leading non-competitor that does not hold stock $j$. $DS\ HHI_{j,t}$ is the HHI of data scientists for stock $j$. $w_{i,j,t}$ is a portfolio weight of stock $j$ in investor $i$'s portfolio. |

Table A2. ONET Codes of Data Scientists

| ONET Code | Occupation | Data Scientists Category |
|---|---|---|
| 15-2051.00 | Data Scientists | Data Analytics |
| 15-2041.00 | Statisticians | Data Analytics |
| 15-1299.06 | Digital Forensics Analysts | Data Analytics |
| 15-2051.01 | Business Intelligence Analysts | Data Analytics |
| 15-2051.02 | Clinical Data Managers | Data Analytics |
| 15-1242.00 | Database Administrators | Data Collection |
| 15-1243.00 | Database Architects | Data Collection |
| 15-1212.00 | Information Security Analysts | Data Maintenance |
| 15-1243.01 | Data Warehousing Specialists | Data Maintenance |
| 15-1299.05 | Information Security Engineers | Data Maintenance |
| 15-1299.04 | Penetration Testers | Data Maintenance |

Table A3. Leading 50 Institutional Investors by Data Scientist Employment

This table lists the top 50 institutional investors ranked by the total number of data scientists employed in 2021.

| Investor | NumDS | NumAnalysis | NumCollect | NumMaintain |
|---|---|---|---|---|
| Morgan Stanley & Co. LLC | 4311 | 2746 | 259 | 1306 |
| Credit Suisse Asset Management, LLC (US) | 3288 | 1877 | 245 | 1166 |
| Goldman Sachs & Company, Inc. | 3076 | 2213 | 95 | 768 |
| Liberty Mutual Insurance Group | 2353 | 1535 | 126 | 692 |
| Blackrock Alternatives Management, LLC | 1820 | 1484 | 52 | 284 |
| Fidelity | 685 | 468 | 48 | 169 |
| Bank of the West | 617 | 293 | 19 | 305 |
| LPL Financial LLC | 389 | 203 | 25 | 161 |
| TD Securities, Inc. | 352 | 282 | 7 | 63 |
| Brown Brothers Harriman & Company | 325 | 186 | 29 | 110 |
| Wells Fargo | 309 | 242 | 10 | 57 |
| Kemper Corporation | 252 | 153 | 17 | 82 |
| Susquehanna International Group, LLP | 237 | 115 | 33 | 89 |
| First Bancorp, Inc | 230 | 87 | 27 | 116 |
| Protective Life Corporation | 205 | 148 | 10 | 47 |
| PIMCO (US) | 196 | 102 | 17 | 77 |
| Barclays Capital Inc. | 179 | 129 | 13 | 37 |
| United Bank | 178 | 77 | 34 | 67 |
| First National Bank of Omaha | 157 | 81 | 14 | 62 |
| StoneX Group Inc. | 155 | 69 | 11 | 75 |
| CNO Financial Group, Inc. | 142 | 78 | 2 | 62 |
| MFS Investment Management | 136 | 88 | 5 | 43 |
| Fisher Investments | 134 | 85 | 12 | 37 |
| BMO Capital Markets (US) | 133 | 98 | 9 | 26 |
| Bridgewater Associates, LP | 133 | 47 | 1 | 85 |
| abrdn Inc. | 128 | 80 | 3 | 45 |
| Balyasny Asset Management LP | 118 | 67 | 6 | 45 |
| Federated Hermes MDTA LLC | 118 | 72 | 9 | 37 |
| BNP Paribas Asset Management USA, Inc. | 115 | 65 | 8 | 42 |
| Allianz Global Investors U.S. LLC | 104 | 68 | 6 | 30 |
| Barings LLC | 102 | 42 | 9 | 51 |
| State Street Global Advisors (US) | 102 | 60 | 9 | 33 |
| Citizens Financial Group, Inc. | 96 | 51 | 9 | 36 |
| AmTrust Financial Services, Inc. | 94 | 59 | 7 | 28 |
| JP Morgan Asset Management | 92 | 78 | 2 | 12 |
| Ally Financial Inc. | 89 | 38 | 10 | 41 |
| RBC Global Asset Management (U.S.) Inc. | 79 | 56 | 6 | 17 |
| AQR Capital Management, LLC | 77 | 54 | 4 | 19 |

| Investor | NumDS | NumAnalysis | NumCollect | NumMaintain |
|---|---|---|---|---|
| Neuberger Berman, LLC | 75 | 57 | 1 | 17 |
| Bank of Oklahoma, N.A. | 72 | 41 | 16 | 15 |
| Millennium Management LLC | 72 | 26 | 6 | 40 |
| APG Asset Management US, Inc. | 67 | 21 | 5 | 41 |
| Tower Research Capital LLC | 67 | 29 | 8 | 30 |
| Aegon Asset Management US | 66 | 40 | 1 | 25 |
| Bill & Melinda Gates Foundation | 65 | 49 | 3 | 13 |
| Dimensional Fund Advisors, L.P. | 64 | 39 | 14 | 11 |
| Commonwealth Financial Network | 62 | 31 | 4 | 27 |
| Bayview Asset Management, LLC | 61 | 25 | 9 | 27 |
| South State Bank | 56 | 30 | 2 | 24 |
| The Vanguard Group, Inc. | 56 | 16 | 6 | 34 |