Asset (and Data) Managers

Marco Zanotti*

Swiss Finance Institute, USI Lugano

Job Market Paper

January 25, 2026

Abstract

This paper studies whether asset management companies use customer data to attract capital. Exploiting information from their websites' codes, I track when fund managers begin collecting and analyzing data on their potential customers using tools like Google Analytics or A/B testing. I show that funds adopting such technologies attract 1.5% higher annual flows and charge higher fees, despite no improvement in performance. These results are concentrated in retail share classes and decline with competition as more rival funds adopt similar tools. At the fund-family level, adopters expand their product offerings, and new funds focus more on retail-oriented themes. Within existing funds, I find evidence of changes in prospectus content and greater sales efforts rather than product differentiation. Overall, data technologies allow managers to raise more capital and charge higher fees, without passing these monetary gains on to investors. These findings show that technological innovation in asset management extends beyond portfolio allocation decisions, and it affects how funds attract and retain capital.

Keywords: Asset Management, Data Economy, Information, FinTech
JEL Classification: G14, G23, L10

*Swiss Finance Institute, USI Lugano. Email: marco.zanotti@usi.ch. I am very grateful to Francesco Franzoni, Laurent Frésard, Alberto Plazzi, and Andrea Tamoni for their continuous guidance and support. I also thank Simona Abis, Nolwenn Allaire, Federico Baldi-Lanfranchi, Maxime Bonelli, Francesco D'Acunto, Silvia Dalla Fontana, Rich Evans, Thierry Foucault, Ahmed Guecioueur, Anton Lines, Roxana Mihet, Rachel Nam, Alessandro Previtero, Jonathan Reuter, Steven Riddiough, Alberto Rossi, Stefano Rossi, Vatsala Shreeti, Rafael Zambrana, Anthony Lee Zhang as well as seminar and conference participants at WFA, 7^{th} Future of Financial Information, Annual Financial Markets and Liquidity Conference, SGF Conference, Trans-Atlantic Doctoral Conference, 32^{nd} Finance Forum, HEC Paris PhD Workshop, FMARC Doctoral Tutorial, SFI Research Days, Amsterdam Business School (UvA), CU Boulder, Dartmouth Tuck, HEC Montréal, Pompeu Fabra, Bank of Italy, Stockholm School of Economics, Toronto Rotman, UCSD Rady SOM, Verona, Warwick Business School and USI Lugano for insightful comments and discussions. I acknowledge financial support from INQUIRE Europe.

1 Introduction

Asset management companies use large volumes of data to inform investment decisions. They often process rich datasets to predict asset payoffs and guide their portfolio choices. Yet, asset managers compete for capital as much as they compete for alpha. A significant share of their revenues—and employees' compensation—is tied to the total assets under management (AUM) and the inflows they attract.[1] As a result, their incentives are largely aligned towards attracting flows and increasing the total assets they manage. This makes understanding investor demand as important as predicting asset payoffs. However, little is known about whether and how asset management companies use data to improve their ability to attract capital, and how these practices impact the structure of the industry.

This lack of evidence is surprising, given that firms across the economy increasingly collect detailed information from their interactions with customers. Such data allow companies to refine products, target marketing, and predict demand, but they also raise important questions on whether the rents from data are shared between firms and consumers—e.g., through lower product prices, or better product quality.

The asset management industry provides a clean setting to study these questions. Fund companies operate in a regulatory framework that requires disclosure of product characteristics (holdings), prices (fees), and measures of product quality such as performance —information rarely available in other product markets. In this market, investors are a fund's *customers*. They range from unsophisticated retail clients to large institutions, allowing to observe heterogeneous responses to an increase in managers' data availability.

In this paper, I show that asset managers use customer data to improve their ability to attract and retain capital. I exploit information from their websites' codes and track the adoption of technologies used to analyze customer data, such as Google Analytics or A/B testing tools. These technologies, which I refer to as *data technologies*, allow to gather rich information about customers—such as demographics, browsing patterns, and product interest—that managers can use to tailor communication and product design. I find that funds adopting data technologies attract about 1.5% more inflows per year and charge higher fees, despite showing no improvement in performance. The effects are concentrated in retail share classes and, within funds, arise from targeting better potential investors. At the company level, adopters tend to launch more funds in retail-oriented categories, consistent with product differentiation at the margin. Altogether, these results show that asset managers extract value from analyzing customer data and retain all the monetary rents it generates.

[1] See Cen, Dou, Kogan and Wu (2024) for details on US mutual funds' contract structure. Ibert, Kaniel, Van Nieuwerburgh and Vestman (2017) find similar compensation structure in Swedish mutual funds.

In canonical models of the asset management industry, funds' customers (i.e., investors) learn managers' skills over time (Berk and Green, 2004). They observe risk-adjusted performance, update their prior, and reallocate capital accordingly. The implicit common assumption in the literature has been that this learning is one-sided—investors learn about managers, but managers have nothing to learn about investors. In practice, however, funds have increasingly access to customers' information. For example, Broadridge, a leading FinTech company, offers asset managers tools that identify and analyze website visitors by matching their IP addresses to geographical locations. By linking these data to the headquarters of large investors such as pension funds, managers identify prospect clients and reach out with targeted offers. Broadridge also provides analytics on future demand for asset management products, information on the most effective distribution channels (e.g., financial advisors vs. direct-to-consumer), and several analyses on both own and competitors' investors. Understanding whether the increasing availability of data affects how managers attract AUM is important. If access to customer data changes how funds raise capital, it will also influence how investors allocate their savings, with potential implications for wealth distribution and financial markets. In this paper, I show that access to customer data materially affects the equilibrium of the asset management industry.

My empirical strategy relies on information from asset managers' websites to quantify their willingness to collect and analyze investors' data. A key challenge in answering whether asset managers use customer data to attract flows is that we do not observe which managers adopt such practices. To overcome this limitation, I track when asset managers install technologies designed to collect and analyze customers' data. Every website is made of different building blocks, called *technologies*. For example, e-commerce often install technologies enabling secure payments, such as Apple Pay or PayPal. I identify technologies designed to collect and analyze customers' data, such as Google Analytics, and the exact installation date on asset managers' websites. These *data technologies* allow to store digitized information about user interactions on website pages. For instance, they enable A/B testing (tools similar to Randomized Control Trials), identify which channels bring more potential customers, track users' search history, demographics, and audience overlap with competitors. By observing the adoption date of data technologies by asset managers, I can proxy for when a given manager begins collecting investors' information. This approach provides a novel measure to proxy for the adoption of data technologies that are otherwise unobservable.

I use a staggered difference-in-differences framework to study how asset managers change after installing technologies that process customers' data. I compare changes in monthly flows (pre- and post-adoption) within the same fund and month, controlling for several commonly used covariates. This approach allows me to reduce concerns that time-invariant

fund characteristics or common style shocks are driving the results. First, I find that asset managers receive more flows after adopting a data technology on their website. Funds using technologies that analyze investors' information receive approximately 1.5% more flows each year. The magnitude of this effect is not negligible: it is comparable to the increase in flows associated with a fund delivering an annual alpha of 1.75% (the 84th percentile in my sample). In other words, the effect is similar in size to the flow-performance sensitivity for a fund at the 84th percentile of the performance distribution.[2] Measuring inflows and outflows separately, the result originates from an increase in inflows with no significant change in outflows, implying that funds attract new capital after adoption.

Second, funds charge higher expense ratios following the adoption of data technologies. The effect on inflows raises the question of whether asset management companies use the same information advantage to adjust their pricing. I show that managers with more customer data charge higher fees. This result shows that technological innovation does not always reduce search cost frictions.[3] New technologies, such as smartphone apps, might reduce investors' search costs, as they allow easier access to information at their fingertips. These results, however, reveal a more complex effect of technology on the asset management industry. When technology gives managers —not investors— more information, fees rise rather than fall.

I conduct a series of tests to verify that asset managers extract useful information from customer data. Mutual fund data at the share class level offer a clear laboratory for studying the heterogeneity of the effects between retail and institutional investors. I find that the effects are concentrated in retail share classes, with no impact on institutional share classes *within* the same fund. These results are consistent with managers learning from customer data, as the technologies I study in this paper are more informative about retail investors.

The effects are unique to website technologies designed to collect and analyze customers' data. When I conduct placebo tests using plugins unrelated to web traffic data, such as Google Maps, I find no results. Furthermore, consistent with a learning channel, the value of data declines as more competitors observe similar information: the effects weaken as more funds within a fund category adopt these technologies. On a similar note, data technologies have decreasing marginal returns, a common feature in information goods (Veldkamp, 2011). Additionally, after the adoption of a data technology, funds maintain a smaller cash buffer and hold more illiquid assets, consistent with managers facing less uncertainty about

[2]The flow-performance sensitivity is the positive relationship between a fund's risk-adjusted performance and fund flows (Chevalier and Ellison, 1997).

[3]Technology is commonly viewed as reducing search frictions. For example, from Basten and Ongena (2020) "search costs are lowered when lending moves to the type of online platform we study". See, among many others, Thakor (2020); Argyle et al. (2022) for examples in which FinTech innovations reduce search costs.

investors' liquidity needs.

The above results do not warrant a causal interpretation, as the adoption of data technologies is a choice made by asset managers. Even though I find no evidence of pre-treatment trends, managers may install these tools when they anticipate higher demand for their products. To alleviate endogeneity concerns, I use two plausibly exogenous sources of variation. First, because the local geographical environment plays an important role in technology diffusion (Moretti, 2004; Conley and Udry, 2010; Gennaioli et al., 2013), I instrument adoption with the local supply of graduates in data analytics-related fields. I obtain the number of bachelor's, master's, and Ph.D. graduates from U.S. universities in data analytics, statistics, and computer science from the Integrated Postsecondary Education Data System (IPEDS). I then divide the total number of graduates in each fund's commuting zone by its population, and use this measure as an instrument for a fund's decision to adopt a data analytics technology. Intuitively, the instrument captures variation in local expertise that lowers the cost of adoption. Because completing a degree requires several years, any local shock that increases enrollment today would affect graduate supply only with a lag, reducing concerns that contemporaneous shocks drive both the instrument and fund outcomes. The IV estimates confirm the main results: plausibly exogenous adoption of a data technology raises annual flows by 1.7% and increases expense ratio. These effects are not driven by large financial districts such as Boston, Chicago, New York, Los Angeles, Philadelphia, or San Francisco (Christoffersen and Sarkissian, 2009).

Second, I exploit variation in the information that funds can extract from web traffic data, which is plausibly unrelated to fund flows. Specifically, I use the public release of TensorFlow, a widely used open-source machine learning (ML) library in November 2015.[4] TensorFlow release improves the precision of prediction models in settings with large data availability. For example, Uber and Airbnb integrated TensorFlow to develop their ML algorithms for rider-driver matching and pricing models, among other things. Intuitively, if the widespread availability of ML allows to extract more informative signals from data, the impact of data technology adoption should intensify after TensorFlow's release. Consistent with this conjecture, the impact of data technologies is about 30% larger post-release. To tighten identification and validate this mechanism, I exploit cross-sectional heterogeneity in data availability prior to the release of TensorFlow. I construct two proxies for the size of each fund's dataset at the release date: (i) the number of months between adoption and November 2015, and (ii) the number of distinct data technologies installed. These measures capture the amount of data available to a fund when TensorFlow is released (i.e., (i) proxy for the time series of customer data collected up to TensorFlow's release, and (ii) proxy for

[4]For reference on the release, see: wired.com/[...]/google-open-sources-its-artificial-intelligence-engine.

the cross-sectional size of the dataset). Using a difference-in-differences specification with (i) and (ii) as continuous treatment intensity, I find that funds with larger datasets benefit more from TensorFlow's release. Importantly, these results do not require that all funds actively use ML to analyze customer data; this setting is similar to an intent-to-treat (ITT) design.

I next examine *how* funds benefit from data. Managers gain from data primarily through two channels: by targeting existing products more effectively, and by differentiating their product offerings to cater to investor demand. Within existing funds, access to customers' data improves marketing and sales to prospective clients (Roussanov et al., 2020; Chen et al., 2022). To test this mechanism, I compare active and passive funds. Passive funds have little discretion in portfolio holdings, which limits their ability to differentiate through product design. If results were originating solely from product differentiation, I should observe no effect on passive funds. Instead, I find similar effects across active and passive funds, consistent with a targeting mechanism. To provide more direct evidence for this, I use data from SEC N-SAR filings to show that adopting funds increase in-house sales and marketing expenditures, and reallocate 12b-1 payments away from external broker-dealers toward their own captive retail sales force. Fund prospectuses also become easier to read and more directly oriented toward retail investors. These changes suggest that funds use data to better target investors rather than to persuade or obfuscate (Mullainathan et al., 2008). I find no significant change in product differentiation within fund, suggesting that the main effect on existing funds stems from cosmetic changes.

At the fund-family level, the product differentiation mechanism plays a complementary role (Massa, 2003; Loseto and Mainardi, 2023; Bonelli et al., 2023). Fund families that adopt data technologies are better able to anticipate demand for new themes and adjust their product menus accordingly. They launch more new funds than non-adopters, and those funds focus more on themes that attract retail interest, such as ESG, artificial intelligence, and cybersecurity (Ben-David, Franzoni, Kim and Moussawi, 2023).

I consider several alternative explanations for my findings. One rationale behind the results might be that funds adoption of new technologies correlates with a superior ability to generate performance. Although I control for past performance across all specifications and the results on retail share classes (with no effect on institutional share classes within fund) are inconsistent with this hypothesis, I formally test and reject this conjecture using different measures of risk-adjusted performance. Another plausible explanation is that fund managers are not learning about investor preferences but rather persuading them. For instance, adoption could coincide with a rebranding of the fund. I show that this explanation is unlikely in two ways. First, suppose the results reflected funds' rebranding. In that case, they should also hold for other website plugins that are unrelated to data collection, as it

is likely that websites rebranding update several plugins simultaneously. However, I find no results using placebo technologies that are not used to collect web traffic data. Second, after adopting data technologies, fund prospectuses become more readable and transparent, inconsistent with obfuscation or persuasion motives (Mullainathan et al., 2008; Ellison and Ellison, 2009; deHaan et al., 2021).

Overall, my findings show that the impact of new technologies in the asset management industry extends beyond asset allocation. By learning about investors rather than securities, managers earn rents by better matching products to investor demand rather than by delivering alpha. Greater access to investors' data might affect incentives to acquire information and generate performance, with implications for financial markets. Moreover, managers retain all the monetary gains from data. Understanding the origin of the additional capital to the industry is important to assessing whether new technologies are net beneficial. For example, if these flows come from shifting capital from passive investments to thematic funds, the welfare implications may be ambiguous. More broadly, identifying if similar mechanisms operate in other services, product markets, or industries is key to understanding the broad economic role of data. Whether data grow the pie or redistribute existing rents is a central question for the new data economy.

Related Literature. This paper contributes to three main strands of literature. First, it adds to the growing literature on the role of new technologies in finance (Abis and Veldkamp, 2023). Existing research studies how technology affects financial forecasting (e.g., Chi, Hwang and Zheng, 2024; Coleman et al., 2022; van Binsbergen, Han and Lopez-Lira, 2022; Dessaint, Foucault and Frésard, 2024), stock market quality (e.g., Farboodi and Veldkamp, 2020; Martin and Nagel, 2022; Dugast and Foucault, 2024), households (e.g., Mihet, 2022; D'Acunto and Rossi, 2023; Rossi and Utkus, 2024) and capital allocation (Abis, 2022; Bonelli, 2024; Birru et al., 2024; Bonelli and Foucault, 2024; Sheng et al., 2025). While most of this evidence focuses on how technology influences asset allocation, I show that it also directly affects fund managers' ability to attract and retain capital. In doing so, this paper highlights a direct link between technological innovation and the equilibrium allocation of capital.

Second, the paper contributes to the literature on the industrial organization of the asset management industry.[5] Hortaçsu and Syverson (2004) show that investors' search costs are central to explaining why homogeneous S&P500 index funds charge different fees. When investors face search frictions, they choose from a limited subset of available products rather than the best option overall, generating price dispersion even in homogeneous markets.

[5]Gârleanu and Pedersen (2018) link the efficiency of asset prices to the efficiency of the market for asset management services. Their findings bridge the (in)efficiencies in both markets and emphasize the importance of the asset management industry for asset prices.

Roussanov, Ruan and Wei (2020) highlight the importance of mutual funds' marketing for attracting investors' capital[6] (Reuter and Zitzewitz, 2006; Kostovetsky and Manconi, 2018; Chen et al., 2022). More recently, Obizhaeva (2024) finds that ETFs attract more flows when advertising through online search engines. According to this strand of literature, technological improvement will reduce search frictions and increase competition among managers. This paper shows a more nuanced role of technology: when information advantages accrue to managers instead of investors, fees rise rather than falling. A related line of literature studies asset managers' strategic product market choices (e.g., Massa, 2003; Kostovetsky and Warner, 2020; Cvitanić and Hugonnier, 2022; Betermier et al., 2023; Loseto and Mainardi, 2023; Bonelli et al., 2023). In these works, fund families change their product menus to reduce investors' switching costs or to differentiate themselves. To the best of my knowledge, this paper is the first to investigate whether asset managers collect investors' information to better meet their demand. In traditional rational models (e.g., Berk and Green, 2004; Berk and van Binsbergen, 2015), learning is one-sided: investors learn about managers' skill by observing their performance over time. This paper complements these frameworks by showing that learning also operates in the opposite direction —managers learn from investors' data. In that sense, this work bridges the literature on asset management to studies on customer capital (Gourio and Rudanko, 2014; Belo et al., 2014), defined as the stock of a firm's customer relationship. Roldan-Blanco and Gilbukh (2021) relate firms' customer capital to markup dynamics[7], while He, Mostrom and Sufi (2024) show that publicly listed U.S. firms invest more than 4% of their revenue on sales and marketing to build customer capital. This paper shows that advances in data technologies strengthen asset management companies' ability to build and monetize from customer capital. The same logic may apply to other industries as well.

Third, this paper contributes to the literature on the role of data in the economy (e.g., Brynjolfsson and McElheran, 2016; Goldfarb and Tucker, 2019; Jones and Tonetti, 2020; Cong et al., 2021). Chung and Veldkamp (2024) review this growing literature in detail. The central insight is that the increasing amount of digitized information is valuable for economic agents, and a no-data equilibrium differs from a data economy (Farboodi and Veldkamp, 2023). While data are not conceptually distinct from information, what differs is the enormous number of data points available and the sources from which agents extract that information. A crucial question in understanding the role of data in the economy is whether it increases utilitarian welfare or redistributes rents (Baley and Veldkamp, 2025). I

[6]This literature intersects with research studying what drives investors' flows to asset managers and their effect on asset prices (e.g., Dou, Kogan and Wu, 2024). See Christoffersen et al. (2014) for a survey.

[7]Complementary works on this growing literature include Morlacco and Zeke (2021), Baker et al. (2023), and Arellano-Bover et al. (2025).

show that funds generate more value added (Berk and van Binsbergen, 2015) after adopting data technologies, and that this additional value originates from retail investors rather than institutional investors. This distinction is crucial for understanding where the surplus from data arises and who captures it. Under strict neoclassical assumptions, the adoption of data technologies is associated with larger utilitarian welfare, as it improves matching between investors and managers. However, if these assumptions do not hold, for example, because of households' biases, data may redistribute rents from retail investors to fund managers. Overall, I show that valuable data for asset managers are not only datasets for identifying investment opportunities (Farboodi et al., 2021, 2024; Bonelli and Foucault, 2024), but also customers' data.

2 Theoretical Framework

In this section, I describe the main hypotheses regarding the role of data analytics technology for asset managers' ability to attract capital. I summarize the main intuition here and develop a simple economic framework in Appendix A. The theoretical framework builds on Berk and Green (2004) and Berk and van Binsbergen (2015). In these traditional models of competitive markets for asset management services, investors learn about fund managers' skills by observing their risk-adjusted performance over time. However, the learning is one-sided. Whether managers do or do not have information about investor preferences, tastes, or values is irrelevant for the equilibrium of the industry.

I maintain a competitive rational framework and extend it to a setting where managers compete not only on risk-adjusted performance, but also on how well they understand customer preferences. I will use the terms *investor* and *customer* interchangeably. The goal is to understand whether information about investors' preferences plays a role in the allocation of capital to asset managers.

As in Berk and van Binsbergen (2015), investors allocate wealth across a continuum of funds based on risk and return. In my setting, investors may also have non-pecuniary preferences. These attributes might include a preference for holding green assets (e.g., warm-glow preferences, Hartzmark and Sussman, 2019; Pástor, Stambaugh and Taylor, 2022), or a specific taste over the fund's communication style and frequency[8] (e.g., daily push notifications, monthly emails, etc.). The precise mix of such preferences evolves slowly over time. Investors know their own tastes, but fund managers do not observe them directly.

[8]This interpretation is reminiscent of the effect of "trust" in Gennaioli, Shleifer and Vishny (2015). Previtero and Xing (2025) also model non-pecuniary preferences in the mutual fund sector. They study the value added by financial advisors in a competitive market.

Instead, managers infer preferences from observing noisy signals.

The key friction is that investors' non-pecuniary preferences are latent, and asset managers differ in how precisely they can learn them. Some managers have better data, better survey infrastructure, or more sophisticated analytics. Others operate with more noisy information. This asymmetry creates variation in how well different managers learn about shifting investors' demand. Fund managers accumulate information over time, building their stock of knowledge. This stock of knowledge corresponds to "data" in traditional models of the data economy (Farboodi and Veldkamp, 2023; Abis and Veldkamp, 2023).

Managers who know more about investors' taste can better anticipate shifts in demand and tailor their offering accordingly. For example, if younger investors increasingly prefer digital communication, an informed manager could introduce mobile apps or chatbots early and attract more capital than competitors with noisier signals.[9] This mechanism leads to the first hypothesis I test:

Hypothesis 1. (Flows and Data Technologies)
Asset managers collecting more information about investor preferences receive larger fund flows.

The additional flows are the result of a better match between funds and investors, beyond the performance they deliver. This logic extends to other product preferences. For instance, if investors' demand for AI or green assets intensifies, more informed managers can forecast the shift and adjust their product offerings accordingly. As a consequence, they can attract capital even without superior performance (Ben-David, Franzoni, Kim and Moussawi, 2023). Less informed competitors may miss the trend and lose assets to funds that offer better alignment with investors' tastes.

The theoretical framework also predicts differences in the equilibrium fund fees. Asset managers who better align their products with investors' taste can charge higher fees. These managers are harder to replace for investors, since outside options offer a worse fit. They are willing to pay more, even if these products might deliver lower performance. Returning to the example above, a fund that caters to green preferences may be preferable to a generic alternative even if it generates lower performance. Thus, my second testable hypothesis concerns fees, i.e., expense ratio:

[9] In a recent interview, Arpit Sarin, Vice President of Digital at Regions Bank, mentioned "understanding how GenZ and TikTok investors pick their investment" and "how much do they trust financial advisors" as urgent questions for which an asset manager "would need a lot of data to get an answer" (aiandbanking.libsyn.com/[...]/ai-driven-digital-transformation-in-banking).

Hypothesis 2. (Expense Ratio and Data Technologies)
Funds with more information about investor preferences charge higher fees compared to competitors.

When a fund's product offering aligns more closely with investor preferences, those consumers are willing to pay more for the product. Selling more specialized products allows funds to charge higher fees in equilibrium. This prediction is reminiscent of common features in IO models (e.g., Lancaster, 1966; Salop, 1979; Pellegrino, 2024 among others).

In sum, this framework emphasizes that information about investors' preferences, tastes, or values might matter for the equilibrium of the asset management industry. Once we introduce a role for non-pecuniary preferences in investors' utility, information about customers' demand allows managers to attract more assets and charge higher fees. This holds even in competitive and rational markets. In the following sections, I test these and other ancillary predictions using data technologies used to track customer behavior as a proxy for the willingness to analyze investors' data.

3 Data and Measurement

The data I use in this paper come from multiple sources. This section describes the datasets and summarizes the main cleaning steps; full details are in Appendix B. First, I introduce the main sample of US mutual funds and ETFs. Second, I complement these data with information from N-SAR regulatory filings, mutual fund prospectuses, and portfolio holdings. Finally, I describe website technology data and document a series of facts regarding the adoption of data technologies in asset management.

3.1 Data

My main data sources are the CRSP Survivorship Bias-Free Mutual Funds dataset, Morningstar Direct, and FactSet Funds. I merge CRSP with FactSet by CUSIP, which identifies unique financial securities and is not re-assigned over time. Then, I merge with Morningstar using CUSIP and ticker. I follow Berk and van Binsbergen (2015) as closely as possible in cleaning funds' data. I outline the main steps here and provide thorough details in Appendices B.1 to B.4.

The sample consists of US equity mutual funds and ETFs. I adjust all AUM numbers by inflation (in January 2000 dollars) and remove observations prior to a fund's first offer date to mitigate incubation bias concerns (Evans, 2010). I exclude funds with less than two years in the sample and those whose (inflation-adjusted) AUM never exceeds $5 million

(Kacperczyk, Sialm and Zheng, 2008). Funds data are available at the share class level; that is, different share classes belonging to the same fund are reported separately. Therefore, for each month, I aggregate share classes at the fund level by summing the AUM of all subclasses and weighting all other variables (e.g., fees, returns) by lagged AUM. I identify fund families following Dannhauser and Spilker (2023), and estimate fund alphas using 24-month rolling regressions on monthly returns. The sample is from March 1993 to December 2023. The beginning date is March 1993 because earlier AUM data are limited (Pástor, Stambaugh and Taylor, 2015). The final sample comprises 8,125 funds (7,836 equity mutual funds and 289 ETFs), with 947,079 fund-month observations. My sample is somewhat larger than comparable samples of US mutual funds in the literature. The reason is twofold. First, incorporating FactSet yields a more comprehensive merge with CRSP than using Morningstar alone. Second, I do not remove index funds, institutional share classes, sector funds, or funds that allocate less than 80% of their portfolio to stocks.

Following prior literature (e.g., Lou, 2012), I compute the investment flow to fund i in month t as

$$Flow_{i,t} = \frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t}) - MRG_{i,t}}{AUM_{i,t-1}} \times 100, \tag{1}$$

where $AUM_{i,t}$ is the assets under management (total net assets) of fund i in month t, $r_{i,t}$ is the monthly (gross) return, and $MRG_{i,t}$ is the increase in AUM due to fund's mergers happening in month t. Accounting for $MRG_{i,t}$, I avoid misattributing funds' mergers as inflows. I winsorize all variables at the 1% and 99% levels.

Table 1 about here

Table 1 presents summary statistics for all fund-month observations in my sample. The distribution of AUM is rightly skewed, as is common in the institutional investors' literature. The average expense ratio is 1.12%, with 0.28 percentage points attributable to marketing and distribution expenses (12b-1 fees). On average, funds in my sample have 0.12% negative flows each month and negative net abnormal returns (after fees), consistent with evidence for the U.S.

I complement the main funds' data with three additional sources: N-SAR filings, mutual fund prospectuses, and portfolio holdings.

N-SAR filings. Until June 2018, investment management companies filed semi-annual N-SAR Forms with the SEC.[10] Each Form N-SAR reported fund-level information, but was submitted by the fund complex (i.e., the SEC registrant). A fund complex may include

[10]Starting in June 2018, the SEC replaced N-SAR with Forms N-CEN and N-PORT.

multiple funds. I link all funds within a fund complex to the main CRSP–Morningstar–FactSet dataset (Appendix B.7 provides a detailed description of the steps). N-SAR filings contain information not available in traditional datasets, including redemption fees, marketing expenses, captive retail sales forces, 12b-1 fee rebates, and separate reporting of sales (inflows) and redemptions (outflows).[11]

Fund Prospectuses. Mutual fund prospectuses are from the SEC's EDGAR (Forms 485APOS and 485BPOS). I closely follow Abis (2022) and Mullally and Rossi (2025) to retrieve prospectuses from EDGAR. The sample period for the prospectus data spans from 2006 to 2023. This sample begins in 2006 because in that year, the SEC started requiring series IDs (fund identifiers) and class IDs (share class identifiers) in their filings. These identifiers enable a direct link to the main sample. For each filing, I parse the text, strip HTML tags, and extract the Principal Investment Strategy (PIS) section (Abis, 2022; Abis and Lines, 2024), see Appendix B.8 for details on the procedure.

Portfolio Holdings. Finally, mutual fund portfolio holdings are from Thomson Reuters (s12). I use quarterly equity holdings from 2004Q2 to 2023Q4. The holdings' sample starts in 2004Q2 because the SEC began requiring quarterly holdings disclosure only in May 2004 (Rule 30b1-5).

3.2 Data Technologies

I identify asset managers' willingness to collect data using information from their websites.

In particular, I rely on tags contained in the website's source code to detect whether managers adopt plugins designed to analyze web traffic data. Such plugins are widely used in several industries to track customer behavior and preferences. For example, Amazon and Nike, in their cookie policy, mention analyzing web traffic data to "improve their products" (Amazon, 2025) and to "understand personal preferences" (Nike, 2025).

Asset managers are increasingly using these plugins to learn about investor behavior and adapt their products. Vanguard, for example, reports using Adobe Analytics to identify which content is more likely to drive engagement for specific visitors, allowing to target investors more effectively. Similarly, Reliance Mutual Fund tracks web traffic to analyze investors' behavior and test alternative website layouts. This practice is typically called A/B testing, and resembles randomized control trials (RCTs) on website pages to run personalization experiments.[12] These plugins provide a natural setting for studying questions about the role

[11] For other papers using information from N-SAR filings see, among others, Edelen (1999); Chernenko and Sunderam (2020); Evans, Gomez and Zambrana (2024).

[12] NIMF Mutual Funds Chief Digital Officer, Arpanarghya Saha, notes that "we need to make our products and services easy to understand —similar to approaches taken at more traditional e-commerce companies."

of customer data in the asset management industry.

All websites are made by different tools that work as building blocks. These building blocks are typically referred to as *technologies*. For instance, installing Google Maps technology allows a website to display an interactive map on its pages (e.g., to show store locations). Other technologies, such as Adobe Analytics, are designed to collect and analyze web visitors' data. I obtain information on websites' technology adoption from BuiltWith, an alternative data provider. BuiltWith analyzes websites' source code and searches for specific patterns, such as HTML tags, that identify the presence of technologies.[13] They continuously crawl websites to capture installed technologies, starting January 2000. As a result, I observe the exact month a website installs (and eventually removes) a given technology. Henceforth, I define *data technologies* as those aimed at collecting and analyzing customer information, such as Google Analytics or Adobe Analytics.

I merge BuiltWith data with fund websites in the CRSP-Morningstar-FactSet sample. Fund website information is primarily from CRSP. However, CRSP starts consistently reporting each fund's website only in January 2008. To extend coverage, I hand-collect all website registration dates from `whois.com` and back-fill each fund's website from its registration date to December 2007.[14]

Figure 1 and Table 2 about here

Figure 1 plots the adoption of data technologies among US funds. The blue area shows the total number of funds with at least one data technology on their website in a given month. The red line reports the share of funds adopting data technologies. A few funds began adopting data technologies as early as 2006. However, the significant surge occurred in 2012, when the adoption rose from 15% to 30%. This surge coincides with a major release of Google Analytics, which remains the leading provider in the space. The adoption of data technologies in asset management continued to grow after 2012 and stabilized around 2019. As of December 2023, approximately 70% of the funds in the sample (over 2,000 unique funds) had adopted at least one data technology.

In Table 2 I list the leading data technologies, by the end-of-sample. Google Analytics accounts for the lion's share of adoption, with around 60% of funds installing it.[15] Other

The firm highlights data technologies as one of its major recent priorities (business.adobe.com/[...]/mutual-fund-case-study).

[13] See also Charoenwong et al. (2024) for another usage of BuiltWith's data in finance.

[14] This procedure marginally increases the number of technologies' adoptions, as most data technologies spread after 2011 (see Figure 1). However, this cleaning step ensures I don't misclassify funds as non-adopters before January 2008.

[15] Appendix E.2 shows that all the main results are not driven solely by Google Analytics (Appendix Table E.9).

leading technologies include Omniture Test & Target, which enables A/B testing[16], as well as LiveRamp, which integrates big data across platforms.

All technologies in Table 2 are designed to generate signals from web visitors' preferences. These features are not limited to the most commonly used data technologies by asset managers in my sample. Appendix Figure E.2 shows the word cloud built from descriptions of all data technologies installed by asset managers. Common terms include "tracking", "insight", and "analytics". Overall, these technologies suggest that asset managers actively track customers to inform their decisions. The remainder of the paper examines the implications of these technologies for adopters and for the equilibrium of the asset management industry.

A concern with using this proxy is that adoption might be decided by hosting providers, rather than the funds themselves. For instance, AWS Web Hosting may pre-install the same set of technologies by default across all the websites it hosts. If that happens, then the actual adoption rates would not reflect the choices made by the funds. To ensure this is not the case, I collect data on the hosting provider used by all websites in my sample. For each group of websites hosted by a hosting provider in a given month, I compute the average cosine similarity in technology adoption. The cosine similarity measures the overlap between two vectors. Intuitively, if two websites using the same hosting provider in a given month share the same set of technologies, the cosine similarity equals one. When the cosine similarity is lower than one, the websites installed different technologies within hosting-month group.

Appendix Figure E.1 presents the distribution of cosine similarities within each hosting-month group. The mean similarity is 0.29, with a median of 0.27, both of which are far from one. Websites sharing the same hosting provider install markedly different technologies.[17] Thus, it is unlikely that website technology adoption is determined by the hosting services.

4 Main Findings

In this section, I examine how the adoption of data technologies affects asset managers. Guided by the framework in Section 2, I first study the impact of data technology on fund flows and fees (expense ratio). I conclude this section with a series of robustness tests in support of the results.

[16] A/B tests are tools similar to RCTs, which are increasingly used in several industries. They allow to randomly split the web traffic audience and study several alternatives of product bundles, pricing, or descriptions.

[17] For comparison, Panel B of Appendix Figure E.1 show results from a simulation in which adoption is random within groups. The cosine similarity in the actual data is lower than what would be expected under random assignment.

4.1 Funds Flows and Data Technologies

The first main hypothesis (Hypothesis 1, Section 2) predicts that funds with more data about investor preferences attract more flows. I test this prediction by studying whether funds receive more capital after adopting a data analytics technology on their website. For each fund i and month t, I define a dummy variable $\text{DATA}_{i,t}$ equal to one if the fund has at least one data technology in place. Then, I estimate the following specification:

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta \, \text{DATA}_{i,t} + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \tag{2}$$

where $Flow_{i,t+1}$ is the fund i's flow between months t and $t+1$, defined as in equation (1). α_i and η_t are fund and time fixed effects, respectively. I also include a specification with fund category×time fixed effects, to mitigate concerns about category-specific shocks. $\boldsymbol{X}_{i,t}$ is a set of control variables, including fund size ($logAUM$), performance, (log) age, turnover, and 12b-1 fees.[18] I measure performance using the CAPM alpha, as previous research shows it is the closest asset pricing model to what mutual fund investors use (Berk and van Binsbergen, 2016; Barber, Huang and Odean, 2016). Using alternative measures, such as Fama–French 3-factors, 5-factors, or alpha with respect to all available Vanguard index funds (Berk and van Binsbergen, 2015) yields the same conclusions (see Appendix E).

The coefficient of interest is θ. Importantly, including fund and time fixed effects implies that identification comes from variation in flows before versus after data-adoption, relative to the same change for funds without data technology. I account for the staggered treatment design in my difference-in-differences estimations (Callaway and Sant'Anna, 2021; Gardner et al., 2024). I cluster all standard errors at the fund and month levels.

Table 3 about here

Table 3 shows that data adoption leads to larger fund flows. The first row reports the coefficient of interest, θ. Columns (1)–(2) present baseline difference-in-differences (OLS) results, while columns (3)–(4) show estimates corrected by staggered difference-in-differences concerns (following Gardner et al., 2024). Columns (2) and (4) add fund category×time fixed effects to absorb category-specific shocks. I omit coefficients on control variables for brevity.[19]

[18] Expense ratio and lagged flows are common controls in fund flows regressions. Yet, both variables are affected by the treatment. As a consequence, including these variables as controls could introduce post-treatment bias. For completeness, I report results with these additional covariates in Appendix Table E.4. Results are qualitatively unchanged.

[19] All controls enter with the expected sign across all specifications, e.g., positive flow-performance sensitivity (Chevalier and Ellison, 1997; Sirri and Tufano, 1998; Pástor et al., 2015; Franzoni and Schmalz, 2017). For comparison with previous works, Appendix Table E.1 shows detailed results.

Across all specifications, funds that adopt a data technology on their website experience significantly larger inflows after adoption. The coefficients imply about 0.13% (1.56%) larger monthly (yearly) inflows for data-driven funds after adoption. These effects are economically significant. For example, the effect is comparable to the flow-performance sensitivity (Chevalier and Ellison, 1997) associated with a fund generating a 1.75% annual alpha (84^{th} percentile in my sample).

I report results for both baseline difference-in-differences and estimators addressing heterogeneous treatment effects or other concerns in staggered designs (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). The results are similar across all approaches. This is not by chance. In my setting, the total weight attached to "forbidden comparisons" (Goodman-Bacon, 2021) is less than 8%. This relatively low weight arises because the main sample begins several months before the first adoption (treatment), which allows precise estimation of group effects (Gardner, 2021; Gardner et al., 2024). A similar logic applies to period (time) effects. These results remain unaffected by controlling for funds' Morningstar Rating (Ben-David, Li, Rossi and Song, 2022), using alternative sets of control variables, or removing funds that adopt only Google Analytics. In Section 4.4, I thoroughly discuss robustness tests.

In Appendix E.2, I examine the parallel trends assumption implicitly behind the identification in equation (2), and covariates balance tests. While the parallel trend assumption is, by definition, untestable, I find no evidence of statistically significant differences in pre-adoption trends between treated and controls (Table E.2). I also cannot reject the hypothesis of covariates balance before adoption at conventional confidence levels (Table E.3, Figure E.3).

Figure 2 shows the dynamics of the effect on flows. I interact the coefficient of interest with event-time dummies in a stacked sample, following Gormley and Matsa (2011). Then, I report monthly event-time coefficients up to one year before/after adoption, with 95% confidence intervals. The estimates show no significant pre-trends before adoption. The treatment effect is persistently positive only after three months since adoption.

Figure 2 about here

Moreover, since multiple funds in a given fund family may share the same website (on average, a family has 1.30 websites), a potential concern is that θ could reflect within-family effects. I address this by showing robustness to: (i) clustering standard errors at the *fund family* and month levels, (ii) restricting only to families with one website per fund, and (iii) aggregating observations within family-month.

The fund flows discussed above (equation (1)) represent *net* flows. This measure captures the difference between inflows and outflows for fund i during month t. To better understand the impact of data technology on fund flows, I investigate whether the results stem from larger inflows or smaller outflows. I use data from the SEC's N-SAR filings, which required funds to report their monthly inflows and outflows separately.[20] The N-SAR sample spans from January 2006 to June 2018, after which the SEC replaced N-SAR with N-CEN and N-PORT filings. Table 4 shows the results for fund inflows and outflows, separately.

Table 4 about here

Columns (1) and (2) report estimates on inflows, while columns (3) and (4) on outflows. The table shows that the effect of data technology on fund flows arises through larger inflows. The first two columns resonate with previous results: funds that install data technology attract significantly more inflows after adoption. By contrast, the effect on outflows is statistically indistinguishable from zero. Therefore, after adoption, funds receive larger inflows, with no significant difference in outflows. This result suggests that the effect operates mainly on the extensive margin rather than the intensive margin (i.e., funds attracting new investors). This is important because it points to new technologies allowing asset managers to collect more (*new*) capital from investors.

Asset managers ultimately care about scale –that is, the total AUM. While data adoption attracts more flows, there might be some conditions under which this does not translate into larger scale. The increase in flows may be small in dollar terms, short-lived, or offset by weak performance.[21] To directly verify that adoption increases fund size, I estimate the same specification using the (log) AUM as dependent variable –excluding lagged size from the control variables set to avoid post-treatment bias (Roberts and Whited, 2013). Appendix Table E.5 report the results. In line with the results on flows, funds that adopt data technologies manage about 12% larger AUM after adoption.

Taken together, these results support the view that data technologies help asset managers attract more capital. Funds that adopt data technology are associated with larger flows after adoption. This result stems from more inflows, without affecting outflows.

4.2 Increase in Expense Ratio

The second main hypothesis (Hypothesis 2, Section 2) predicts that funds with more customer data charge higher expense ratios in equilibrium. This happens because fund

[20] I refer to Appendix B.7 for further details on N-SAR filings data.

[21] Later in the paper, I show that none of these cases is supported empirically.

managers align better with investor preferences, beyond the alpha the fund delivers to them. For example, if these technologies allow managers to offer funds closer to investors' taste or values, they can elicit a higher willingness-to-pay and increase fees.

Table 5 about here

I study this hypothesis with a similar specification to equation (2), where the expense ratio (in %) is on the left-hand side. Table 5 reports the results. The estimates imply that funds using data technology increase their fees by approximately 2 to 4 basis points after adoption (first row). Results are similar including category×time fixed effects, and/or other control variables such as the Morningstar rating. Further, the increase in expense ratio is independent of changes in marketing and distribution fees, as I do not find significant changes in 12b-1 fees.

A legitimate concern might be that this increase in fees reflects a mere composition effect within fund. The expense ratio of funds with both retail and institutional share classes is the AUM-weighted average expense ratio across share classes. Because retail share classes are more expensive, a fund's expense ratio could rise mechanically if retail share classes attract more capital than institutional ones. I rule out this explanation, running the specification for a sample of funds with only retail share classes (Appendix Table E.14).

Theory of the industrial organization of asset management predicts that new technologies will reduce fees by lowering investors' search costs (e.g., Hortaçsu and Syverson, 2004; Gârleanu and Pedersen, 2018; Roussanov et al., 2020 among others). Technological development makes it easier for investors to find and compare asset managers. With easier search, investors contact more managers, identify better funds, and drive the market toward the first-best with no price dispersion and fees equal to marginal cost.[22] Thus, these models imply that technological innovations will tend to reduce the cost of financial services. The prediction from Section 2 is markedly different and offers a different perspective on the role of new technologies for the equilibrium of the asset management industry. This difference arises because the data technologies I study in this paper facilitate information acquisition for asset managers, not customers. When managers —not investors— benefit from technological innovation, fees rise in equilibrium rather than fall.[23]

[22]For example, from Roussanov et al. (2020): "With the advancement in information technology and the emergence of services enabling more transparent comparison between funds, we would expect the search frictions to decline over time."

[23]Relatedly, Buchak, Chau and Jørring (2023) document that FinTech lending raises borrowing costs in the U.S. mortgage market.

4.3 Endogeneity Concerns

The adoption of data technology may be correlated with unobservable factors. For example, fund managers may install data analytics tools just before an expected surge in product demand. I address this concern using two sources of variation that are plausibly exogenous to fund flows and fees. First, I instrument adoption with the local supply of data analytics graduates. Second, I exploit the release of TensorFlow, a widely used machine learning library that improved the precision of signals managers can extract from existing datasets.

4.3.1 Instrumental Variable: Data Analytics Graduates

It is well known that the local environment plays a role in the diffusion of technology. Regions with specialized human capital foster adoption through local complementarities and knowledge spillovers. For example, Moretti (2004) and Gennaioli et al. (2013) show that local human capital spurs technology adoption by firms. On a similar note, Conley and Udry (2010) document that local learning externalities accelerate adoption even in developing economies.

The local environment also plays a significant role in asset management. Christoffersen and Sarkissian (2009) show that funds in large cities outperform because managers in financial districts interact more frequently and share investment ideas (Cujean, 2020). I test whether this intuition extends to technology adoption in asset management. Appendix Table E.15 shows that funds are more likely to adopt when nearby funds do. To separate geographical effects from peer effects, I regress a fund's adoption decision on the (lagged) share of adopters in its style and in its geographic location. If adoption were driven by within-style peers, the lagged share of adopters in a fund's style would predict adoption. Instead, only the share of adopters in the same zip code, city, or state matters, consistent with a local environment effect. The magnitude of this local effect declines as the geographic unit expands —from zip code to city to state— as in agglomeration economies with technology diffusion (Rosenthal and Strange, 2004).

When I allow endogenous adoption of a data technology in my economic framework, the prediction echoes intuition from the literature on technology diffusion: Local concentration of experts in data analytics facilitates diffusion by lowering the marginal cost of adoption. Proximity to specialized human capital facilitates learning and access to technology.[24]

[24] Hoberg and Neretina (2024) apply a similar intuition from agglomeration economics to instrument for a firm's participation in trade associations. They use the likelihood that a firm learns about a trade association via social connections.

Guided by this idea, I construct an instrument based on the local supply of data analytics graduates. Specifically, I use the number of university graduates in data analytics-related fields from universities near a fund's headquarter. I obtain the annual number of bachelor's, master's, and Ph.D. degrees awarded in data analytics, statistics, and computer science by U.S. universities from the Integrated Postsecondary Education Data System (IPEDS).[25] For each year, I aggregate the total number of graduates within a U.S. commuting zone (CBSA) and scale by the commuting zone's population. This variable proxies for local concentration of experts that can implement data analysis.

Because the number of graduates varies at the annual frequency, I aggregate flows at the fund-year level and measure expense ratios annually. This avoids mismatches in frequency that could create aggregation bias or mismeasurement (Ghysels et al., 2006). Then, I instrument a fund's adoption choice with the local supply of data analytics graduates.

The exclusion restriction requires that local data analytics graduates affect fund outcomes only through adoption. This assumption is plausible because fund fixed effects absorb all time-invariant differences across funds, including location choice. A reasonable concern is that local shocks could simultaneously increase graduate supply and investor demand. However, this concern is mitigated by the timing of graduation: completing a degree requires between two to four years, so any local shock that raises enrollment in university programs today would affect graduate supply only with a delay, whereas shocks to investor demand would likely materialize faster. I further address this point scaling the instrument by the CBSA's population, so that it captures talent supply rather than market size. Table 6 shows the results.

Table 6 about here

I report first-stage results in columns (5) and (6). The first stage confirms that funds in areas with more data analytics graduates are more likely to adopt. An additional data analytics graduate per 100 people predicts 1.1% higher probability of installing tools aimed at tracking customers' data. This effect is sizeable but plausible: a one-standard deviation increase in local graduates ($\sigma_{IV} = 0.60$) raises adoption by 1.33% ($0.60 \times 1.1\%/0.495 = 1.33\%$) relative to the mean adoption rate of 49.5%. For context, while not quantitatively comparable in outcome, Moretti (2010) documents large local externalities of human capital —one additional skilled job generates roughly +2.5 jobs locally.

The second stage uses variation from local graduates to estimate the effect of adoption on fund flows (columns (1)-(2)) and fees (columns (3)-(4)). The Kleibergen-Paap statistic to

[25] See Appendix C for the detailed list of Core Instructional Programs (CIPs) in these fields.

test for weak instruments is around 20 across all specifications, above the threshold of 10 for rejecting weak instruments (Stock and Yogo, 2005). The second stage results are consistent with earlier evidence. Adoption of a data technology leads to a significant increase in fund flows. At the same time, those funds charge higher expense ratios. On average, plausibly exogenous adoption raises annual flows by 1.7%, (column (2)) and expense ratios by about 3 bps (column (4)). These magnitudes are close to the baseline results in Tables 3 and 5. Overall, the IV estimates quantitatively and qualitatively confirm the main predictions, reducing concerns that results purely reflect endogeneity in the adoption.

The results are not driven by big financial districts such as Boston, Chicago, New York, Los Angeles, Philadelphia, or San Francisco. In Appendix Tables E.16 and E.17, I report robustness results removing these financial centers as well as including CBSA×time fixed effects.

4.3.2 Shock to Information Precision: Open Source Machine Learning

In this section, I use a different source of variation to further address endogeneity concerns. I exploit the release of an open-source machine learning library, called TensorFlow, in November 2015. TensorFlow's public release provides a plausibly exogenous increase in the precision of signals that managers can extract from a given dataset. TensorFlow reduces the cost of training machine learning algorithms in settings where large amounts of data are available. For example, Uber, which collects data through its app, used TensorFlow to predict the probability of a successful customer-driver match.[26]

If machine learning raises prediction precision from a given dataset, the effect of data technologies should be stronger after November 2015. To limit concerns that managers adopted a data technology *because* of TensorFlow's release, I exclude (i) funds that adopt after November 2015 and (ii) funds that adopt within six months before the release.[27] This restriction limits endogeneity concerns in adoption: all treated funds in this sample had a technology in place as of November 2015.

As a first step, I compare the effect of data technologies on flows before and after TensorFlow's release. I interact $\text{DATA}_{i,t}$ with a dummy equal to one post-November 2015 and zero otherwise (denoted $Post_t$):

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 \, \text{DATA}_{i,t} + \theta_2 \left(\text{DATA}_{i,t} \times Post_t\right) + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \tag{3}$$

[26] Several other large platforms such as Airbnb, Kakao Mobility, and Twitter started using TensorFlow soon after its release to improve prediction systems (tensorflow.org/about/case-studies).

[27] Results are unchanged when extending the exclusion window to 12 months.

where θ_2 captures the additional effect of installing a data technology after TensorFlow's release.

Table 7 about here

Columns (1) and (2) in Table 7 report the results. The first row shows estimates for the baseline coefficient on the DATA dummy. Consistent with earlier evidence, unconditionally, funds adopting data technology attract significantly larger flows ($\theta_1 > 0$). The second row reports results for the coefficient of interest in this specification: θ_2. Consistent with machine learning improving prediction precision, the effect of data technologies is about 30% higher after TensorFlow's release.

Next, I tighten identification exploiting cross sectional heterogeneity right before the release of TensorFlow. I construct two continuous treatments that proxy for the *amount* of data available to each fund as of November 2015. The first continuous treatment is the tenure of adoption, i.e., the number of months between a fund's first adoption of a data technology and TensorFlow's release. Intuitively, this proxies for the length of the time-series of data available for training machine learning algorithms. The second continuous treatment is the number of data analytics technologies installed on a fund's website. The idea behind this proxy is that funds with more data technologies can collect more customers' characteristics and, thus, have larger datasets. For example, Adobe Analytics collects demographic information on web visitors, while Hotjar lets users observe heatmaps of customers' interaction —i.e., detailed information on where users click. Then, I estimate:

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 \text{DATA}_{i,t} + \theta_2 \big(\text{DATA}_{i,t} \times Post_t\big) + \theta_3 \big(\text{DATA}_{i,t} \times Post_t \times z_i\big) + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \quad (4)$$

where z_i is either the tenure of adoption as of November 2015, or the number of technologies installed (i.e., the continuous treatments introduced above). The coefficient of interest is the interaction with the continuous treatment: θ_3. In both cases, I expect stronger post-release effects for funds with larger z_i; equivalently, $\theta_3 > 0$.

Columns (3)-(4) and (5)-(6) in Table 7 report results using as continuous treatment the tenure of technology adoption and the number of technologies installed, respectively. The third row shows estimates for the coefficient on continuous treatments. Across all specifications, both continuous treatments confirm the conjecture that the additional effect post-TensorFlow release is stronger for funds with (ex-ante) larger datasets.

Appendix Table E.19 shows that this specification is robust to excluding growth funds. A concern may be that the release of TensorFlow could raise expectations about growth firms'

cash flows, or lower their discount rates, attracting flows to growth funds. For instance, if tech firms benefit more from machine learning than other firms, investors may expect higher future growth for those stocks. The estimates above already include fund category×month fixed effects, which mitigate concerns that category-specific shocks drive the results. Yet, in the Appendix I show that the results remain unchanged when I exclude growth funds.

4.4 Summary of Additional Robustness Results

The results above may still reflect variation other than the funds' tracking of customers' data. In Appendix E, I run a series of robustness tests to address this concern. I summarize these tests here and report results in the Appendix.

First, Ben-David, Li, Rossi and Song (2022) shows that mutual fund investors rely on simple labels, such as Morningstar ratings, when allocating capital. If data technology adoption correlates with ratings, my results might be spuriously capturing those rating changes. I address this concern controlling for Morningstar rating in Appendix Tables E.4 and E.6 (columns (3), (4), (7), and (8)). Results are quantitatively and qualitatively unchanged. This may not appear surprising: Morningstar ratings are assigned based on relative risk-adjusted performance. Thus, mutual funds (and Morningstar) have little discretion over their assignment.

Second, as many fund families host multiple funds on their website (the average is 1.3 websites per family), a reasonable concern might be that the effect comes from a family-specific shock correlated with the adoption (treatment). I address this in three ways. The first approach is similar to how the empirical corporate finance literature deals with treatment at the group-level when multiple firms enter the same group (e.g., multiple firms within a U.S. state in which a law is passed, Giroud and Mueller, 2010; MacKinnon et al., 2023). I cluster standard errors at the *fund family* (group) and time level. In Appendix Tables E.10 and E.11, I show that the results are robust to this conservative clustering. In the second approach, I restrict the sample to fund families with only one website per fund (Appendix Table E.12). Results are unchanged. The third approach is to aggregate observations at the family-month level. Appendix Table E.13 confirms the results hold.

Third, I use alternative measures of risk-adjusted performance beyond the CAPM alpha. Results are robust and quantitatively stable using different measures of alpha (Appendix Tables E.7 and E.8).

Fourth, as expense ratios (and flows) change after adoption, including them as controls could induce post-treatment bias (Roberts and Whited, 2013). For this reason, my baseline specification excludes expense ratio and past fund flows. However, since these controls are

standard in the fund flows literature, I test robustness by adding them in Appendix Tables E.4 and E.6. Results hold.

Finally, the results are not driven by a unique feature in Google Analytics —the most widely adopted technology in my sample (see Table 2). One concern may be that Google prioritizes funds that use Google Analytics in common browser searches, making those funds more visible. I re-estimate the baseline specification excluding Google Analytics from the set of data technologies in Appendix Table E.9. The results remain robust: estimates are not driven by Google Analytics, but reflect a feature common to all data technologies.

5 Additional Effects of Learning from Customers' Data

I interpret the results above as evidence that asset managers extract useful information from data. However, other channels may be consistent with similar evidence. To strengthen this interpretation, I test additional hypotheses and rule out alternative explanations.

I first show that the effect is concentrated only on retail share classes, without affecting institutional ones, consistent with web traffic data being informative about retail customers. Next, I examine competitive effects: If funds learn from data, the effect should diminish as more competing funds adopt similar tools. I then test for diminishing returns to information. According to a learning channel, the marginal benefit of adoption should decline with the number of technologies already in place. Further, I test whether adoption reduces redemption uncertainty for fund managers. If managers indeed learn from data, having more information about customers may allow them to hold a smaller cash buffer, allocate more to illiquid assets, and lower redemption fees. Finally, I rule out alternative explanations such as superior performance after adoption or funds' rebranding.

5.1 A Learning Channel

The findings above hinge on the idea that fund managers learn information from web traffic data. Arguably, web traffic data reveal more information about retail than institutional investors. Thus, the effect of data adoption should be concentrated in retail share classes. Mutual funds data offer a natural setting to test this prediction. I can compare the effect of data technology on retail and institutional share classes *within fund*. I run a specification similar to equation (2), but at the share class level:

$$Flow_{j,i,t+1} = \alpha_i + \eta_t + \theta^I \, \text{DATA}_{j,i,t} + \theta^R \, (\text{DATA}_{j,i,t} \times \text{Retail}_{j,i,t}) + \boldsymbol{\beta}' \boldsymbol{X}_{j,i,t} + \varepsilon_{j,i,t+1}, \quad (5)$$

where $Flow_{j,i,t+1}$ is the flow of share class j of fund i in month t. $\text{Retail}_{j,i,t}$ is a dummy equal to one if the share class j of fund i is sold to retail investors. Including fund fixed effects allows to compare the effect on different share classes within the same fund. The coefficient of interest is θ_R, which measures the incremental effect of data adoption on retail share classes. Learning about retail investors implies a joint prediction: flows of institutional share classes should not respond to data adoption ($\theta^I = 0$), while only retail share classes should ($\theta^R > 0$).

Table 8 about here

Table 8 reports the results and confirms this joint prediction. The effect of adoption is concentrated in retail share classes, without any impact on institutional ones. The coefficient θ^R (second row in columns (1)-(2)), is positive and significant across all specifications. The magnitude aligns with earlier results: data technologies are associated with approximately 2% higher flows per year (0.175% monthly). By contrast, estimates for institutional share classes are statistically indistinguishable from zero (first row in columns (1)-(2)). Estimates are unchanged when running the specification in equation (2) for retail and institutional share classes separately (columns (3)-(4)). These results support the view that website technologies generate signals about retail investors, but provide no information on institutional investors.

I emphasize that this result does not imply that funds cannot learn useful information on institutional investors or that information about those investors is irrelevant. Instead, the evidence shows that this technology —web traffic analytics— does not generate signals about institutional investors' behavior.

The benefit of data also depends on who else has it. If many funds adopt similar data technologies, the marginal value from adoption should decline. This is a common feature of the value of information in settings with strategic substitutes (e.g., Grossman and Stiglitz, 1980). Therefore, I test whether there is a competition effect in the data space. I define a coefficient that captures "data competition" *within* a fund category as:

$$\gamma_{c,t} = \frac{\sum_{i=1}^{N_{c,t}} \text{DATA}_{i,c,t}}{N_{c,t}}, \tag{6}$$

where $N_{c,t}$ is the total number of funds in category c in month t; and the sum in the numerator counts funds with data technology in place, in category c at time t. A larger $\gamma_{c,t}$ implies that a larger fraction of funds within a fund-category collect investors' data. I then estimate specification (2) by bins of $\gamma_{c,t}$ at the time of adoption. Figure 3 presents the results.

Figure 3 about here

First movers benefit significantly more from adopting data technologies than later adopters. The effect declines monotonically with competition. The leftmost bar in Figure 3 shows that funds adopting when competition is low ($\gamma_{c,t} < 25\%$) attract about 5% more annual flows (0.39% monthly). On the other hand, funds adopting when more than 75% of peers already have data show no effect.

Competition also reduces benefits for funds that have already adopted. Appendix Table E.20 reports regressions with an interaction between adoption and $\gamma_{c,t}$. This term captures how post-adoption competition affects the value of data. The coefficient is negative and significant: benefits from data erode as more competing funds adopt. Estimates suggest that when all peers within a fund's category analyze customers' data ($\gamma_{c,t} \to 100\%$), the additional rents are indistinguishable from zero.

Another natural prediction of a learning channel is that the marginal benefit from adding (new) information should decrease (Veldkamp, 2011). Different data technologies generate different signals. For example, Hotjar provides heatmaps of web traffic interactions, while LinkedIn Insights allows to collect demographic data from social media. Although these signals differ, their incremental value should fall if managers learn from data.

I can test this prediction using the number of technologies installed by a given fund. I group the number of technologies into five bins: $K = 1$, $K = 2$, $K = [3, 10]$, $K = [11, 15]$, and $K > 15$, where K denotes the number of technology installed. Then, I plot the bin coefficients relative to $K = 0$ (the baseline) in Figure 4. Each point estimate shows how flows differ for funds in that bin versus funds with no data technology.

Figure 4 about here

Point estimates rise with K, as a fund adopts more data technologies, but at a decreasing rate. I also overlay a concave line fit, $\hat{y} = \alpha + \beta \log(1 + K)$ (solid blue line). This pattern is overall consistent with decreasing marginal benefits from data. In Appendix Table E.21, I also estimate a specification including the number of data technologies and its square. When the squared number of technologies is included, its coefficient enters negatively (although not significantly) in the regression, suggesting again concave returns to data.

Another implication of learning from data is better liquidity management. Mutual fund managers face uncertainty about investors' redemption needs and therefore hold cash buffers to meet unexpected liquidity demands. Because data reduces uncertainty, managers with more customers' data may better predict redemptions. If so, data managers should hold less

cash, particularly when expected returns are high, as holding cash is most costly in such states.

To test this prediction, I study whether (after adoption) funds reduce their cash buffers when expected returns are high. As my sample includes only equity funds, I can proxy for expected returns using the fund portfolio's dividend–price ratio. Specifically, I run the following regression:

$$w_{i,q}(cash) = \alpha_i + \eta_q + \lambda_1 \cdot D/P_{i,q} + \lambda_2 \cdot (D/P_{i,q} \times \text{DATA}_{i,q}) + \boldsymbol{\beta}' \boldsymbol{X}_{i,q} + \varepsilon_{i,q}, \tag{7}$$

where $w_{i,q}(cash)$ is fund i's portfolio weight in cash (in %) in quarter q, and $\boldsymbol{X}_{i,q}$ is the same fund-quarter controls' vector as in earlier specifications.

Expected returns are high when the dividend–price ratio is high, as equity prices are relatively low compared to dividends. I therefore expect $\lambda_1 < 0$: fund managers, on average, reduce cash to invest in attractive opportunities. The coefficient of interest in this specification is the interaction term (λ_2), which measures whether data managers reduce cash buffers more aggressively when expected returns are high. As before, identification here comes from time-series variation within funds, not cross sectional differences across funds.

Table 9 about here

Table 9 reports the results. The first two rows present estimates for λ_1 and λ_2.

First of all, column (1) confirms the baseline intuition: when the dividend–price ratio is high (higher expected returns), managers reduce cash holdings. Columns (2) and (3) add the interaction with $\text{DATA}_{i,t}$. As predicted, funds with data technology in place maintain a lower cash buffer ($\lambda_2 < 0$). These results support the interpretation that managers extract information from customers' data.

I next examine whether managers reallocate more into *illiquid* assets. Illiquid stocks offer higher expected returns on average, but they are costly to liquidate if investors require fast redemptions. Funds therefore face a risk-return trade-off in holding such assets (Gómez, Prado and Zambrana, 2024). If data reduce uncertainty about redemptions, managers should tilt more toward illiquid holdings after adoption. Consistent with this intuition, reducing uncertainty about investors' liquidity needs might lead managers to hold more illiquid assets after adoption of a data technology.

Columns (5) and (6) in Table 9 confirm the hypothesis. Fund managers who adopt a data technology increase their exposure to illiquid stocks, as measured by the portfolio's Amihud illiquidity ratio. Although beyond the scope of this paper, these findings may raise concerns

about financial stability, as illiquid assets typically force funds to engage in fire sales when unforeseen shocks occur.

On a similar note, fund managers may be more lenient on redemption fees when they can anticipate redemptions. I can test this last hypothesis directly using N-SAR filings, in which funds disclose semi-annually whether they impose redemption fees to investors liquidating their positions. Columns (7) and (8) in Table 9 show that after adoption, funds are about 7% less likely to charge redemption fees to investors ($-0.017/0.240 = -0.07$). This pattern is consistent with data managers being less uncertain about liquidity needs after adoption. Overall, these results support the idea that fund managers learn from customers' data.

5.2 Alternative Channels

The evidence above confirms the interpretation that fund managers use data technologies to learn about investors. However, a limitation I share with most empirical papers on information/data is that I cannot observe a fund manager's information set directly. However, I can show results consistent with a learning interpretation (above) and rule out alternative explanations that might account for the findings above. Here, I examine two alternative stories and show that they are not supported empirically.

The first alternative explanation is that the adoption of data technologies correlates with managers' ability to generate alpha. For example, managers may adopt data technologies right after acquiring data-analysis expertise. If such skills improve the fund's risk-adjusted performance, investors would notice it and allocate more to that fund. Appendix Tables E.22 test this hypothesis using several performance measures, as well as using the recursive demeaning approach of Pástor, Stambaugh and Taylor (2015). Across all specifications, I find no material change in fund managers' ability to generate alpha after adoption. Therefore, since investors should be able to observe managers' superior ability if this were the main driver of my results, this alternative explanation seems implausible.

A second alternative interpretation of my findings is that adoption coincides with a fund's rebranding. For instance, funds may refresh their websites to restore the perception of their brand for investors, and in the process, install new technologies. If this is the case, my results would purely capture cosmetic changes to websites rather than learning. This channel matters because rebranding could explain inflows or product launches, without implying that managers actually extract information from customers' data.

To test this alternative, I conduct a series of falsification tests using placebo websites' plugins that are unrelated to customers' data collection. These include, for example, plugins used to improve a website's performance, as well as feeds (used to publish blogs), or

advertising technologies. I restrict attention to placebo technologies with adoption rates similar to those of data technologies in my sample, to ensure a fair comparison and that results are not being driven by a lack of power. Then, I re-estimate the baseline regression, replacing $\text{DATA}_{i,t}$ with each of the placebo technologies.

Figure 5 about here

Figure 5 reports the 95% confidence intervals for all coefficients on placebo technologies. All estimates are statistically indistinguishable from zero. This result contrasts with the positive and significant effects of data technologies. This evidence rules out a rebranding story and reinforces the interpretation that the main findings reflect managers' learning from customers' data.

Overall, the evidence supports a channel based on fund managers using new technologies to learn about customers and tailor products accordingly. That said, it is important to acknowledge that I cannot entirely rule out the possibility that other mechanisms may explain my results. Any alternative explanation, however, must also fit the broader set of results. For example, alternative interpretations should explain why adoption leads to more illiquid holdings, why the effect is concentrated in retail share classes, and why the results are specific to technologies that collect and analyze customers' data.

6 How do Funds Attract More Flows?

How do funds benefit from data? Asset managers can cater to investors' demands through two main approaches: by selling their existing products more effectively, or by differentiating the products they offer. The first mechanism works by improving targeting and distribution to improve sales. For example, funds might reallocate resources to affiliated bank advisors or employ wholesalers, whose task is to pitch the fund through their distribution channels. The second approach implies modifying the product itself, such as adjusting portfolio choices, to better align with investors' tastes. A common example is rebalancing toward ESG stocks in response to demand for sustainable investing. The results show that both mechanisms play a role. Within existing fund, the benefit comes primarily by becoming better at selling, while at the fund family I find evidence consistent with horizontal differentiation.

6.1 Marketing Better Existing Funds

I first examine the mechanisms at play within existing fund. I compare the effects of data adoption on active and passive funds. Passive funds have little scope for discretionary changes in their portfolio holdings. Therefore, if the effects originate mainly from product differentiation, the results should be concentrated only in active funds. I split the sample into active and passive funds and estimate the main specification separately for each group. I report the results in Table 10.

Table 10 about here

Pure product differentiation is unlikely to be the only driver of how funds benefit from data technology. I find similarly strong results on flows and fees, for both active and passive funds. The first four columns in Table 10 show results on fund flows: columns (1)-(2) for active funds, and columns (3)-(4) for passive funds. In both groups, funds receive significantly more capital after adoption. Columns (5) to (8) confirm the same pattern for the impact of data technologies on fees. All these results are consistent with a selling channel. If product differentiation were the primary mechanism, I should find weak or no effects among passive funds. By analyzing investors' data, funds improve their targeting and selling strategies rather than solely focusing on product design. This resonates with anecdotal evidence. For instance, Broadridge (2023) describes how fund families can use data to identify which distribution channels are more effective and reallocate resources accordingly.

To further support this interpretation, I test whether marketing becomes more effective in attracting flows after the adoption of data technologies. If customer data are informative about potential investors' preferences, marketing and sales efforts might become more targeted and productive. In this case, I expect asset management companies to shift more resources toward sales and marketing, and for those expenditures to attract more flows. I test these predictions using fund-level data on marketing and sales expenditures from N-SAR filings.

N-SAR filings contain detailed information on funds' marketing activities. In particular, I observe total marketing expenditures as well as the allocation of 12b-1 payments across distribution channels. 12b-1 fees are typically referred to as marketing and distribution fees. They represent annual charges that funds deduct from the AUM to finance distribution activities.[28] N-SAR filings report information on how funds distribute their 12b-1 fees payments to different channels, including captive retail sales force, which refers to the *internal*

[28] 12b-1 fees are capped at 1% of AUM annually and are typically reported as part of the fund's expense ratio. Therefore, investors ultimately bear the cost of these fees.

sales force of the fund. Payments to unaffiliated intermediaries must be reported separately. This distinction allows me to distinguish between in-house selling resources from external intermediation.

A salient example of captive retail sales force is J.P. Morgan Asset Management, which distributes its mutual funds through advisors and bankers located in retail branches. These employees are not independent intermediaries but rather staff of the affiliated bank. Compensation for their selling activities is in part financed by 12b-1 fees. A similar example is in-house wholesalers, whose primary role is to pitch funds to financial advisors.[29] Unlike third-party intermediaries, these employees are dedicated exclusively to promoting the firm's own products, and their salaries are internal sales expenditures.

Tables 11 and 12 about here

I divide the total dollar amount paid to captive retail sales force by the fund's AUM. Then, I examine whether funds modify their 12b-1 fees payments after adoption. Columns (1)-(2) in Table 11 show the results. After adoption, managers increase their 12b-1 payments to internal sales resources. This shift toward internal sales is consistent with a selling mechanism that relies more on in-house sales personnel (Kostovetsky and Manconi, 2018).

Appendix Table E.23 shows how managers reallocate 12b-1 payments across categories. Managers increase spending on internal sales forces by cutting payments to brokers and dealers (external). At the same time, funds redirect resources toward printing and mailing prospectuses to prospective investors (column (4)).

I find similar evidence on broader marketing and sales expenditures. These expenditures differ from the 12b-1 allocations because they cover the total costs of sales, rather than just the portion financed by 12b-1 fees. In practice, funds often spend more than what they collect charging 12b-1 fees, supplementing these with other resources. In columns (3)-(4) of Table 11, I use again data from N-SAR filings to examine the change in total marketing expenditures after adoption. Similar to earlier results, total sales expenditures also increase after adoption.

Next, I test whether marketing efforts become more effective after funds adopt data technologies. Roussanov, Ruan and Wei (2020) highlight the importance of mutual funds' marketing for attracting investors' capital (see also Reuter and Zitzewitz, 2006; Kostovetsky and Manconi, 2018; Chen et al., 2022 for similar evidence). If data improve the targeting of marketing efforts, I expect a stronger association between marketing expenditures and

[29] In Appendix Figure E.6 I examine whether results are concentrated in no-load funds, or they also hold for intermediated (broker-sold) funds (Del Guercio and Reuter, 2014). Results are similar for both types of funds.

subsequent fund flows after adoption. Table 12 confirms this prediction. The baseline coefficients in columns (1)–(2) (first row) show that marketing and sales spending, on average, positively predicts future flows: a 1% increase in marketing leads to additional 0.19% flows. In columns (3)–(4), I report results for the interaction term of marketing and sales with the adoption of data analytics technology (second row). The relationship between marketing expenditures and flows strengthens after adoption. The interaction term is statistically and economically significant, consistent with marketing and sales efforts becoming more productive after collecting potential customer data.

If funds cater to retail investors, I should find observable changes in their marketing and distribution material. I test this hypothesis by analyzing whether text in fund prospectuses becomes more appealing to retail investors after adoption. I focus on the Principal Investment Strategy (PIS) section —the key standardized disclosure of a fund's investment strategy (Abis, 2022). I measure the prospectus readability with the Flesch Reading Ease (FRE) index[30] (Flesch, 1948), average words per sentence (deHaan et al., 2021), and by the frequency of second-person pronouns (e.g., "you", "your"), which prior work in marketing links to stronger retail engagement (Cruz et al., 2017). Table 13 presents the results.

Table 13 about here

After adoption, fund prospectuses become easier to read, shorter, and more direct. The readability of prospectuses (FRE index) increases by 3.5% (0.781/22 = 3.55%). At the same time, average sentence length falls by 4%. Prior work interprets longer sentences as evidence of obfuscation (deHaan et al., 2021), as fund managers often use complexity to justify higher fees. In contrast, I find that fund managers *simplify* disclosure once they start collecting data about investors' preferences.

This result is important. The literature on product complexity/obfuscation shows that intermediaries often use complexity to persuade investors and extract rents; here, I find the opposite.[31] These results help distinguish whether managers use data to meet investors' demand or to persuade them more effectively. The evidence points to a favorable effect of data: Table 13 suggests that managers use data to cater to investors' needs (e.g., Gennaioli, Shleifer and Vishny, 2015), rather than to persuade or obfuscate (Célérier and Vallée, 2017).

[30]The FRE index indicates how difficult a passage in English is to understand. A higher score means the text is easier to read.

[31]Starting with Mullainathan et al. (2008), Carlin (2009), and Lerner and Tufano (2011), a growing body of work documents this mechanism. For example, Célérier and Vallée (2017) shows that European banks employed complex descriptions to sell structured products, while Vokata (2021, 2025) find that products marketed to households with complex language offer attractive yields but deliver strongly negative risk-adjusted performance. Genaro et al. (2025) provides a recent overview of this literature.

Fund prospectuses also mention more themes (Ben-David et al., 2023). I measure the frequency at which the prospectus' text includes words linked to themes such as clean energy, cybersecurity, or cannabis (see Appendix D for the entire list of themes and the respective words). Columns (7)-(8) in Table 13 show that these changes are again consistent with an effort to tailor toward retail investors' interests.

I then investigate further the horizontal differentiation mechanism, within fund. Since the two mechanisms are not mutually exclusive, it is still possible that this channel also plays a role. According to this mechanism, funds modify their existing products. For instance, they may rebalance their portfolios towards stocks with characteristics appealing to retail investors. To do so, I test whether adoption leads to significant portfolio differentiation.

I first compute the holdings' distance measure in Hoberg, Nitin and Prabhala (2017). This measure is based on the pairwise distance between fund holdings, and it is similar in spirit to Hoberg and Phillips (2016). I denote the average pairwise distance between fund i and its peers in quarter q, as $\bar{d}_{i,t}$. A higher distance indicates that fund i is differentiating more from other funds, meaning it is more unique.

Table 14 about here

I examine the effects of data technology adoption on product differentiation in Table 14 Estimates on $\bar{d}$ are positive, but quantitatively small and marginally significant. These results do not allow me to reject the null hypothesis that funds increase product differentiation after adoption. At best, the evidence points to only limited change in product differentiation.

I find similar evidence using the idiosyncratic volatility of a fund's portfolio returns. The intuition for this measure is simple. If a fund differentiates more, its return should become more idiosyncratic, as it is less spanned by the market or common risk factors. I compute the idiosyncratic volatility of each stock in a fund's portfolio as the standard deviation of residuals from a Fama-French 3-factors model. Then, I aggregate to construct the portfolio idiosyncratic volatility. I find no significant shift in idiosyncratic volatility after adoption (columns (3)-(4) in Table 14).

I find again similar evidence on a fund's active share (Cremers and Petajisto, 2009). Active share measures the extent to which a fund deviates from its benchmark. It represents the share of holdings that differ from a fund's benchmark and serves as a standard proxy for how far a portfolio deviates from its benchmark. Columns (5) and (6), in Table 14, report no significant change in a fund's active share after adoption. If anything, the effect is mildly negative.[32]

[32]This slight decline is consistent with the prediction of Berk and Green (2004); Berk and van Binsbergen

Finally, Kostovetsky and Warner (2020) argue that funds perceived by investors as more unique face lower flow-performance sensitivity. Because investors view those products as less substitutable, they are less likely to reallocate capital after observing realization of risk-adjusted performance. Consistent with this idea, in Appendix Table E.24 I find that after adopting data technologies, funds exhibit a decline in flow-performance sensitivity, suggesting that investors perceive these funds as more unique.

All the results above rest on changes for an existing fund, as variation in my identification strategy comes from within fund variation. It is plausible that the fund family decides to horizontally differentiate by launching new funds. Therefore, I next examine the mechanism at play within fund family.

6.2 Product Differentiation within Fund Family

Fund families with better knowledge of investors' preferences may expand their product offerings to cater to specific customer demands. To test this prediction, I examine whether fund families increase the number of funds they offer after adopting a data technology, relative to non-adopters. I aggregate observations at the fund family-month level and estimate the following specification:

$$log(\text{N. of Funds}_{f,t+1}) = \alpha_f + \eta_t + \delta \, \text{DATA}_{f,t} + \varepsilon_{f,t+1}, \tag{8}$$

where N. of Funds$_{f,t+1}$ denotes the number of funds offered by family f in month $t+1$. Similarly to previous specifications, the dummy variable DATA$_{f,t}$ equals one if at least one fund within family f has adopted a data technology in month t. Family and time fixed effects ensure that identification comes from variation in the number of funds offered before versus after data adoption, relative to the same change for fund families not adopting data technologies.[33] I report results in columns (1) and (2) of Table 15.

Table 15 about here

Adopting a data technology leads fund families to increase their product offerings. Columns (1) and (2) of Table 15 show that fund families with a data technology offer on average 20% more funds after adoption. In column (2), I control for the fund family's AUM

(2015). When flows push a fund's AUM above the scale that maximizes alpha, their framework predicts that managers index the marginal dollar.

[33]Because the argument of the log in my specification is never zero, the regression does not face the identification challenges that arise when the argument can equal zero (see Chen and Roth, 2023).

and age, as it is plausible that families develop organizational skills, which reduces the cost of setting up a new fund. The results remain unchanged: fund families expand their product menu after adopting a data analytics technology. This result mirrors the effect of other emerging technologies, such as AI, on firms' product portfolios. Babina, Fedyk, He and Hodson (2024) finds that more AI-intensive firms expand their product varieties, as AI facilitates the accumulation of knowledge and reduces uncertainty in innovation.

I next examine whether the newly launched funds cater to specific investor preferences. I collect Principal Investment Strategy (PIS) text from mutual fund prospectuses, and search for words linked to retail-oriented themes such as ESG, AI, cybersecurity, cannabis, political values, space, and video games.[34] Then, I compute the number of such "thematic" words per 100 words of text, and compare newly launched funds by data adopters versus non-adopters. I include year fixed effects to ensure results do not capture the secular rise of thematic funds in recent years (Ben-David, Franzoni, Kim and Moussawi, 2023). Columns (3) and (4) in Table 15 confirm this intuition. Funds launched by families with data technology are more likely to reference themes in high demand from retail investors, mentioning 0.07 more theme-related words per 100 words. This effect amounts to a 45% increase from the sample mean of 0.157 theme-related words every 100 words.

In sum, this evidence indicates that funds attract more flows primarily by selling their existing products better and expanding product offerings. The effects are present for both active and passive funds, and funds increase in-house sales rep expenditures. Prospectuses become clearer and more retail-oriented. At the same time, I find little change in product differentiation within fund. The fund family, on the other hand, launch new funds and meet specific demand.

6.3 Discussion on Value Added

Overall, these findings show that asset managers benefit from customers' data. A natural follow-up question is whether data technologies improve welfare or merely redistribute rents. This is an important question. Recent advances in technologies have increased the ability of many firms to collect detailed information from customers, raising fundamental questions about competitive and welfare consequences of a "data economy".

Whether better matching with investors' preferences improves welfare depends on investors' ex-ante risk exposures —and possibly their privacy concerns. A thorough answer to this question is beyond the scope of this paper, and would require observing detailed household data, such as brokerage accounts. However, under specific assumptions, the

[34] Appendix D lists the complete set of keywords.

mutual fund industry provides a tractable setting to make progress.

Berk and van Binsbergen (2015) develop a measure of value added by fund managers in equilibrium. This measure relies on three assumptions: (i) investors are rational, (ii) markets are competitive, and (iii) managers maximize profits. Under these conditions, a fund's value added equals its assets under management (AUM) multiplied by its expense ratio. According to this measure, in equilibrium, only skilled funds generate high value added because they both attract large AUM and command high fees. This measure allows me to quantify the monetary gains from data technology for the asset management industry. Under the three assumptions above, these monetary amounts map directly into utilitarian welfare gains.

Data technologies raise value added through two channels: larger AUM, and higher fees. All else equal, after adoption, each adopting fund attract approximately 1.5% additional flows annually[35], compounding to approximately +12% in AUM on average (Table E.5). This flow effect alone raises value added by about $600,000 per fund, in January 2000 dollars (about $1.10 million in September 2025 dollars).[36] In addition, funds raise fees by about 3 basis points after adopting data technology. However, this effect is modest, contributing only about $3,000 per year in value added. Therefore, the increase in value added stems almost entirely from larger AUM.

Crucially, funds do not improve risk-adjusted performance after adopting data technology (see Section 5.2). As a result, asset managers capture the entire additional value, rather than sharing it with investors. This surplus represents the monetary value of customers' data accruing to the asset management industry. Across the industry, the additional value added from data technology amounts to approximately $1.5 billion. Under the neoclassical assumptions discussed above, this surplus coincides with utilitarian welfare gains from a better match between products and investors' preferences.

I emphasize that these conclusions are based on strict assumptions. Therefore, this discussion should be interpreted within this context. For example, this logic abstracts from whether additional flows come from reallocating funds away from passive strategies or from shifting consumption into savings. Furthermore, if investors' preferences are sub-optimal, data technology may redistribute welfare rather than create new value. In all these scenarios, broader access to customers' data heightens the importance of educating households on investment decisions.

[35]Fund flows inform about the growth in assets. To evaluate changes in value added, I need to convert these flow effects into their "stock" equivalent (i.e., AUM), reported in Table E.5.

[36]The average fund in my sample adds $5.23 million per year, in January 2000 dollars (about $9.81 million, in September 2025 dollars). This figure is comparable to results in Berk and van Binsbergen (2015), who estimate that funds added about $3.2 million (in January 2000 dollars) annually between 1962 and 2011.

7 Concluding Remarks

The development of new technologies is changing how asset managers operate. While existing research focuses on their impact on portfolio allocation decisions, this paper shows that technological innovation also affects how managers attract and retain capital. Using novel data on website technologies, I show that asset managers actively collect and analyze customers' data, leading to 1.5% higher annual flows for adopting funds.

The effects concentrate in retail share classes, decline with data competition, and improve liquidity management as funds become less uncertain about redemptions. After adoption, managers expand product menus and raise fees, consistent with customers' data helping product placement and pricing strategies. To alleviate endogeneity concerns, I instrument adoption by the local supply of graduates in data analytics and find similar results. Further, I exploit the release of TensorFlow in November 2015, an open-source machine learning library, as plausibly exogenous shock to the precision of signals managers extract from data. These strategies confirm the results that managers benefit from customers' data.

These findings underline the economic importance of asset managers learning investors' preferences to attract capital. Data technologies increase the value added that asset managers extract from financial markets, as funds attract more AUM and charge higher fees. However, managers retain these gains, as the products offered remain largely unchanged.

This mechanism raises important questions about market efficiency and managerial incentives. While better knowledge of investor tastes can help improve matching between funds and investors, it may reduce incentives to generate alpha and shift effort toward targerting and selling. Whether investors' preferences are optimal becomes even more critical for the overall implications of new technologies. Moreover, as managers use investors' data to predict liquidity needs and reduce cash buffers by holding more illiquid assets, technological advances may raise new concerns for financial stability.

References

Abis, Simona (2022) "Man vs. Machine: Quantitative and Discretionary Equity Management," *Working paper*.

Abis, Simona and Anton Lines (2024) "Broken promises, competition, and capital allocation in the mutual fund industry," *Journal of Financial Economics*, Vol. 162, p. 103948.

Abis, Simona and Laura Veldkamp (2023) "The Changing Economics of Knowledge Production," *The Review of Financial Studies*, Vol. 37, No. 1, pp. 89–118, 08.

Amazon (2025) "Amazon Cookie Policy,"Technical report.

Arellano-Bover, Jaime, Carolina Bussotti, Matteo Paradisi, and Liangjie Wu (2025) "The Labor Demand Implications of Brand Capital: Insights from Trademark Transactions in Italy," *Working paper*.

Argyle, Bronson, Taylor Nadauld, and Christopher Palmer (2022) "Real Effects of Search Frictions in Consumer Credit Markets," *The Review of Financial Studies*, Vol. 36, No. 7, pp. 2685–2720, 11.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024) "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, Vol. 151.

Baker, Scott R., Brian Baugh, and Marco Sammon (2023) "Customer Churn and Intangible Capital," *Journal of Political Economy Macroeconomics*, Vol. 1, No. 3, pp. 447–505.

Baley, Isaac and Laura L. Veldkamp (2025) *The Data Economy: Tools and Applications*, Princeton, NJ: Princeton University Press.

Barber, Brad M., Xing Huang, and Terrance Odean (2016) "Which Factors Matter to Investors? Evidence from Mutual Fund Flows," *The Review of Financial Studies*, Vol. 29, No. 10, pp. 2600–2642, October.

Basten, Christoph and Steven Ongena (2020) "The Geography of Mortgage Lending in Times of FinTech," *CEPR Discussion Paper*, No. DP14918.

Belo, Frederico, Xiaoji Lin, and Maria Ana Vitorino (2014) "Brand capital and firm value," *Review of Economic Dynamics*, Vol. 17, No. 1, pp. 150–169.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi (2018) "Do ETFs Increase Volatility?," *The Journal of Finance*, Vol. 73, No. 6, pp. 2471–2535.

Ben-David, Itzhak, Francesco Franzoni, Byungwook Kim, and Rabih Moussawi (2023) "Competition for Attention in the ETF Space," *The Review of Financial Studies*, Vol. 36, No. 3, pp. 987–1042, 08.

Ben-David, Itzhak I., Jiacui Li, Andrea Rossi, and Yang Song (2022) "What Do Mutual Fund Investors Really Care About?" *The Review of Financial Studies*, Vol. 35, No. 4, pp. 1723–1774, April.

Berk, Jonathan B. and Richard C. Green (2004) "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, Vol. 112, No. 6, pp. 1269–1295.

Berk, Jonathan B. and Jules H. van Binsbergen (2015) "Measuring skill in the mutual fund industry," *Journal of Financial Economics*, Vol. 118, No. 1, pp. 1–20.

——— (2016) "Assessing asset pricing models using revealed preference," *Journal of Financial Economics*, Vol. 119, No. 1, pp. 1–23.

Betermier, Sebastien, David Schumacher, and Ali Shahrad (2023) "Mutual Fund Proliferation and Entry Deterrence," *The Review of Asset Pricing Studies*, Vol. 13, No. 4, pp. 784–829.

Birru, Justin, Sinan Gokkaya, Xi Liu, and Stanimir Markov (2024) "Quants and market anomalies," *Journal of Accounting and Economics*, Vol. 78, No. 1.

Bonelli, Maxime (2024) "Data-driven Investors," *Working paper*.

Bonelli, Maxime and Thierry Foucault (2024) "Displaced by Big Data: Evidence from Active Fund Managers," *Working paper*.

Bonelli, Maxime, Anastasia Buyalskaya, and Tianhao Yao (2023) "Information Intermediaries and Financial Product Design: Evidence from Mutual Funds," *Working paper*.

Broadridge (2023) "Market Analytics,"Technical report.

Brynjolfsson, Erik and Kristina McElheran (2016) "The Rapid Adoption of Data-Driven Decision-Making," *American Economic Review*, Vol. 106, No. 5, p. 133–39, May.

Buchak, Greg, Vera Chau, and Adam Jørring (2023) "Integrated Intermediation and Fintech Market Power," *Swiss Finance Institute Research Paper No. 23-67*.

Callaway, Brantly and Pedro H.C. Sant'Anna (2021) "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, Vol. 225, No. 2, pp. 200–230.

Carlin, Bruce I. (2009) "Strategic price complexity in retail financial markets," *Journal of Financial Economics*, Vol. 91, No. 3, pp. 278–287.

Cen, Xiao, Winston Wei Dou, Leonid Kogan, and Wei Wu (2024) "Fund Flows and Income Risk of Fund Managers," *Working paper*.

Charoenwong, Ben, Zachary T. Kowaleski, Alan Kwan, and Andrew G. Sutherland (2024) "RegTech: Technology-driven compliance and its effects on profitability, operations, and market structure," *Journal of Financial Economics*, Vol. 154.

Chen, Jiafeng and Jonathan Roth (2023) "Logs with Zeros? Some Problems and Solutions*," *The Quarterly Journal of Economics*, Vol. 139, No. 2, pp. 891–936, 12.

Chen, Jane, Wenxi Jian, and Mindy Z. Xiaolan (2022) "The Economics of Mutual Fund Marketing," *Working paper*.

Chernenko, Sergey and Adi Sunderam (2020) "Do fire sales create externalities?" *Journal of Financial Economics*, Vol. 135, No. 3, pp. 602–628.

Chevalier, Judith and Glenn Ellison (1997) "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, Vol. 105, No. 6, pp. 1167–1200.

Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng (2024) "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts," *Management Science*.

Christoffersen, Susan E.K. and Sergei Sarkissian (2009) "City size and fund performance," *Journal of Financial Economics*, Vol. 92, No. 2, pp. 252–275.

Christoffersen, Susan, David K. Musto, and Russell Wermers (2014) "Investor Flows to Asset Managers: Causes and Consequences," *Annual Review of Financial Economics*, Vol. 6, No. 1, pp. 289–310.

Chung, Cindy and Laura Veldkamp (2024) "Data and the Aggregate Economy," *Journal of Economic Literature*, Vol. 62, No. 2, p. 458–84, June.

Coleman, Braiden, Kenneth Merkley, and Joseph Pacelli (2022) "Human Versus Machine: A Comparison of Robo-Analyst and Traditional Research Analyst Investment Recommendations," *The Accounting Review*, Vol. 97, No. 5, pp. 221–244.

Cong, Lin W., Danxia Xie, and Longtian Zhang (2021) "Knowledge Accumulation, Privacy, and Growth in a Data Economy."

Conley, Timothy G. and Christopher R. Udry (2010) "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, Vol. 100, No. 1, p. 35–69.

Cremers, K.J. Martijn and Antti Petajisto (2009) "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *The Review of Financial Studies*, Vol. 22, No. 9, pp. 3329–3365.

Cruz, Ryan E., James M. Leonhardt, and Todd Pezzuti (2017) "Second Person Pronouns Enhance Consumer Involvement and Brand Attitude," *Journal of Interactive Marketing*, Vol. 39, pp. 104–116.

Cujean, Julien (2020) "Idea sharing and the performance of mutual funds," *Journal of Financial Economics*, Vol. 135, No. 1, pp. 88–119.

Cvitanić, Jakša and Julien Hugonnier (2022) "Optimal fund menus," *Mathematical Finance*, Vol. 32, No. 2, pp. 455–516.

Célérier, Claire and Boris Vallée (2017) "Catering to Investors Through Security Design: Headline Rate and Complexity," *The Quarterly Journal of Economics*, Vol. 132, No. 3, pp. 1468–1508.

D'Acunto, Francesco and Alberto G. Rossi (2023) "Robo-Advice: Transforming Households into Rational Economic Agents," *Annual Review of Financial Economics*, Vol. 15, No. Volume 15, 2023, pp. 543–563.

Dannhauser, Caitlin D. and Harold D. Spilker (2023) "The Modern Mutual Fund Family," *Journal of Financial Economics*, Vol. 148, No. 1, pp. 1–20.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, Vol. 110, No. 9, p. 2964–96, September.

deHaan, Ed, Yang Song, Chloe Xie, and Christina Zhu (2021) "Obfuscation in mutual funds," *Journal of Accounting and Economics*, Vol. 72, No. 2.

Del Guercio, Diane and Jonathan Reuter (2014) "Mutual Fund Performance and the Incentive to Generate Alpha," *The Journal of Finance*, Vol. 69, No. 4, pp. 1673–1704.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2024) "Does Alternative Data Improve Financial Forecasting? The Horizon Effect," *The Journal of Finance*, Vol. 79, No. 3, pp. 2237–2287.

Dou, Winston Wei, Leonid Kogan, and Wei Wu (2024) "Common Fund Flows: Flow Hedging and Factor Pricing," *The Journal of Finance (forthcoming)*.

Dugast, Jerome and Thierry Foucault (2024) "Equilibrium Data Mining and Data Abundance," *The Journal of Finance*.

Edelen, Roger M. (1999) "Investor flows and the assessed performance of open-end mutual funds," *Journal of Financial Economics*, Vol. 53, No. 3, pp. 439–466.

Ellison, Glenn and Sara Fisher Ellison (2009) "Search, Obfuscation, and Price Elasticities on the Internet," *Econometrica*, Vol. 77, No. 2, pp. 427–452.

Evans, Richard B. (2010) "Mutual Fund Incubation," *The Journal of Finance*, Vol. 65, No. 4, pp. 1581–1611.

Evans, Richard, Juan-Pedro Gomez, and Rafael Zambrana (2024) "MiFID II Research Unbundling: Cross-border Impact on Asset Managers," *Working paper*.

Fama, Eugene F. and Kenneth R. French (1993) "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, Vol. 33, No. 1, pp. 3–56.

Fama, Eugene F and Kenneth R French (2015) "A five-factor asset pricing model," *Journal of Financial Economics*, Vol. 116, No. 1, pp. 1–22.

Farboodi, Maryam and Laura Veldkamp (2020) "Long-Run Growth of Financial Data Technology," *American Economic Review*, Vol. 110, No. 8, p. 2485–2523, August.

——— (2023) "A Model of the Data Economy." NBER Working Paper No. w28427.

Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran (2021) "Where Has All the Data Gone?," *The Review of Financial Studies*, Vol. 35, No. 7, pp. 3101–3138.

Farboodi, Maryam, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran (2024) "Valuing Financial Data," *The Review of Financial Studies*.

Flesch, Rudolf (1948) "A new readability yardstick," *Journal of Applied Psychology*, Vol. 32, No. 3, pp. 221–233.

Franzoni, Francesco and Martin C. Schmalz (2017) "Fund Flows and Market States," *The Review of Financial Studies*, Vol. 30, No. 8, pp. 2621–2673, 03.

Gardner, John (2021) "Two-Stage Differences in Differences," *Working paper*.

Gardner, John, Neil Thakral, Linh T. To, and Yap Luther (2024) "Two-Stage Differences in Differences," *Working paper*.

Genaro, Alan, Jose Maria Liberti, Pedro A. C. Saffi, and Jason Sturgess (2025) "Product Complexity, Investor Experience, and Returns," *Working paper*.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer (2013) "Human Capital and Regional Development," *The Quarterly Journal of Economics*, Vol. 128, No. 1, pp. 105–164.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2015) "Money Doctors," *The Journal of Finance*, Vol. 70, No. 1, pp. 91–114.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov (2006) "Predicting volatility: getting the most out of return data sampled at different frequencies," *Journal of Econometrics*, Vol. 131, No. 1, pp. 59–95.

Giroud, Xavier and Holger M. Mueller (2010) "Does corporate governance matter in competitive industries?" *Journal of Financial Economics*, Vol. 95, No. 3, pp. 312–331.

Goldfarb, Avi and Catherine Tucker (2019) "Digital Economics," *Journal of Economic Literature*, Vol. 57, No. 1, p. 3–43.

Goodman-Bacon, Andrew (2021) "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, Vol. 225, No. 2, pp. 254–277.

Gormley, Todd A. and David A. Matsa (2011) "Growing Out of Trouble? Corporate Responses to Liability Risk," *The Review of Financial Studies*, Vol. 24, No. 8, pp. 2781–2821.

Gourio, François and Leena Rudanko (2014) "Can Intangible Capital Explain Cyclical Movements in the Labor Wedge?" *American Economic Review*, Vol. 104, No. 5, p. 183–88.

Grossman, Sanford J. and Joseph E. Stiglitz (1980) "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, Vol. 70, No. 3, pp. 393–408.

Gârleanu, Nicolae and Lasse Heje Pedersen (2018) "Efficiently Inefficient Markets for Assets and Asset Management," *The Journal of Finance*, Vol. 73, No. 4, pp. 1663–1712.

Gómez, Juan-Pedro, Melissa Porras Prado, and Rafael Zambrana (2024) "Capital Commitment and Performance: The Role of Mutual Fund Charges," *Journal of Financial and Quantitative Analysis*, Vol. 59, No. 2, p. 727–758.

Harris, Lawrence E., Samuel M. Hartzmark, and David H. Solomon (2015) "Juicing the dividend yield: Mutual funds and the demand for dividends," *Journal of Financial Economics*, Vol. 116, No. 3, pp. 433–451.

Hartzmark, Samuel M. and Abigail B. Sussman (2019) "Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows," *The Journal of Finance*, Vol. 74, No. 6, pp. 2789–2837.

He, Bianca, Lauren Mostrom, and Amir Sufi (2024) "Investing in Customer Capital," No. 33171.

Hoberg, Gerard and Ekaterina Neretina (2024) "Do Trade Associations Matter to Corporate Strategies?" *Working paper*.

Hoberg, Gerard and Gordon Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, Vol. 124, No. 5, pp. 1423–1465.

Hoberg, Gerard, Kumar Nitin, and Nagpurnanand Prabhala (2017) "Mutual Fund Competition, Managerial Skill, and Alpha Persistence," *The Review of Financial Studies*, Vol. 31, No. 5, pp. 1896–1929.

Hortaçsu, Ali and Chad Syverson (2004) "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *The Quarterly Journal of Economics*, Vol. 119, No. 2, pp. 403–456.

Ibert, Markus, Ron Kaniel, Stijn Van Nieuwerburgh, and Roine Vestman (2017) "Are Mutual Fund Managers Paid for Investment Skill?," *The Review of Financial Studies*, Vol. 31, No. 2, pp. 715–772.

Jones, Charles I. and Christopher Tonetti (2020) "Nonrivalry and the Economics of Data," *American Economic Review*, Vol. 110, No. 9, pp. 2819–58, September.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng (2008) "Unobserved Actions of Mutual Funds," *The Review of Financial Studies*, Vol. 21, No. 6, p. 2379–2416.

Kostovetsky, Leonard and Alberto Manconi (2018) "How Much Labor Do You Need to Manage Capital?".

Kostovetsky, Leonard and Jerold B. Warner (2020) "Measuring Innovation and Product Differentiation: Evidence from Mutual Funds," *Journal of Finance*, Vol. 75, No. 2, pp. 779–823.

Lancaster, Kevin J. (1966) "A New Approach to Consumer Theory," *The Journal of Political Economy,*, Vol. 74, No. 2, pp. 132–157.

Lerner, Josh and Peter Tufano (2011) "The Consequences of Financial Innovation: A Counterfactual Research Agenda," *Annual Review of Financial Economics*, Vol. 3.

Loseto, Marco and Federico Mainardi (2023) "Oligopolistic Competition, Fund Proliferation, and Asset Prices."

Lou, Dong (2012) "A Flow-Based Explanation for Return Predictability," *The Review of Financial Studies*, Vol. 25, No. 12, pp. 3457–3489.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023) "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, Vol. 232, No. 2, pp. 272–299.

Martin, Ian W.R. and Stefan Nagel (2022) "Market efficiency in the age of big data," *Journal of Financial Economics*, Vol. 145, No. 1, pp. 154–177.

Massa, Massimo (2003) "How do family strategies affect fund performance? When performance-maximization is not the only game in town," *Journal of Financial Economics*, Vol. 67, No. 2, pp. 249–304.

Menzio, Guido (2023) "Optimal Product Design: Implications for Competition and Growth Under Declining Search Frictions," *Econometrica*, Vol. 91, No. 2, pp. 605–639.

Mihet, Roxana (2022) "Financial Information Technology and the Inequality Gap," *Swiss Finance Institute Research Paper*, No. 21-04.

Moretti, Enrico (2004) "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions," *American Economic Review*, Vol. 94, No. 3, p. 656–690.

——— (2010) "Local Multipliers," *American Economic Review*, Vol. 100, No. 2, p. 373–77.

Morlacco, Monica and David Zeke (2021) "Monetary policy, customer capital, and market power," *Journal of Monetary Economics*, Vol. 121, pp. 116–134.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008) "Coarse Thinking and Persuasion*," *The Quarterly Journal of Economics*, Vol. 123, No. 2, pp. 577–619.

Mullally, Kevin and Andrea Rossi (2025) "Moving the Goalposts? Mutual Fund Benchmark Changes and Relative Performance Manipulation," *The Review of Financial Studies*, Vol. 38, No. 4, p. 1067–1119.

Nike (2025) "Nike Cookie Policy,"Technical report.

Obizhaeva, Olga (2024) "Does Search Engine Visibility Help ETFs Attract Flows?" *Working paper*.

Pellegrino, Bruno (2024) "Product Differentiation and Oligopoly: a Network Approach," *American Economic Review (forthcoming)*.

Previtero, Alessandro and Ran Xing (2025) "Beyond Performance: Mutual Funds, Non-Alpha Services, and the Value of Financial Advisors," *Working Paper*.

Pástor, Luboš, Robert F. Stambaugh, and Lucian A. Taylor (2015) "Scale and skill in active management," *Journal of Financial Economics*, Vol. 116, No. 1, pp. 23–45.

——— (2022) "Dissecting green returns," *Journal of Financial Economics*, Vol. 146, No. 2, pp. 403–424.

Reuter, Jonathan and Eric Zitzewitz (2006) "Do Ads Influence Editors? Advertising and Bias in the Financial Media," *The Quarterly Journal of Economics*, Vol. 121, No. 1, pp. 197–227.

Roberts, Michael R. and Toni M. Whited (2013) "Chapter 7 - Endogeneity in Empirical Corporate Finance1," Vol. 2 of Handbook of the Economics of Finance: Elsevier, pp. 493–572.

Roldan-Blanco, Pau and Sonia Gilbukh (2021) "Firm dynamics and pricing under customer capital accumulation," *Journal of Monetary Economics*, Vol. 118, pp. 99–119.

Rosenthal, Stuart S. and William C. Strange (2004) "Evidence on the Nature and Sources of Agglomeration Economies," *Handbook of Regional and Urban Economics*, Vol. 4, pp. 2119–2171.

Rossi, Alberto G. and Stephen Utkus (2024) "The diversification and welfare effects of robo-advising," *Journal of Financial Economics*, Vol. 157.

Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei (2020) "Marketing Mutual Funds," *The Review of Financial Studies*, Vol. 34, No. 6, pp. 3045–3094.

Salop, Steven C. (1979) "Monopolistic Competition with Outside Goods," *The Bell Journal of Economics*, Vol. 10, No. 1, pp. 141–156.

Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan L. Zhang (2025) "Generative AI and Asset Management," *Working paper*.

Shive, Sophie and Hayong Yun (2013) "Are mutual funds sitting ducks?" *Journal of Financial Economics*, Vol. 107, No. 1, pp. 220–237.

Sirri, Erik R. and Peter Tufano (1998) "Costly Search and Mutual Fund Flows," *The Journal of Finance*, Vol. 53, No. 5, pp. 1589–1622.

Stock, James H. and Motohiro Yogo (2005) *Testing for Weak Instruments in Linear IV Regression*, p. 80–108: Cambridge University Press.

Sun, Yang (2021) "Index Fund Entry and Financial Product Market Competition," *Management Science*, Vol. 67, No. 1, pp. 500–523.

Thakor, Anjan V. (2020) "Fintech and banking: What do we know?" *Journal of Financial Intermediation*, Vol. 41.

van Binsbergen, Jules H, Xiao Han, and Alejandro Lopez-Lira (2022) "Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases," *The Review of Financial Studies*, Vol. 36, No. 6, pp. 2361–2396.

Veldkamp, Laura (2011) *Information Choice in Macroeconomics and Finance*: Princeton University Press.

Vokata, Petra (2021) "Engineering lemons," *Journal of Financial Economics*, Vol. 142, No. 2, pp. 737–755.

——— (2025) "Juicing the Coupon Yield: How Banks Extract Rents from Behavioral Biases," *Fisher College of Business Working Paper*, No. 2025-22.

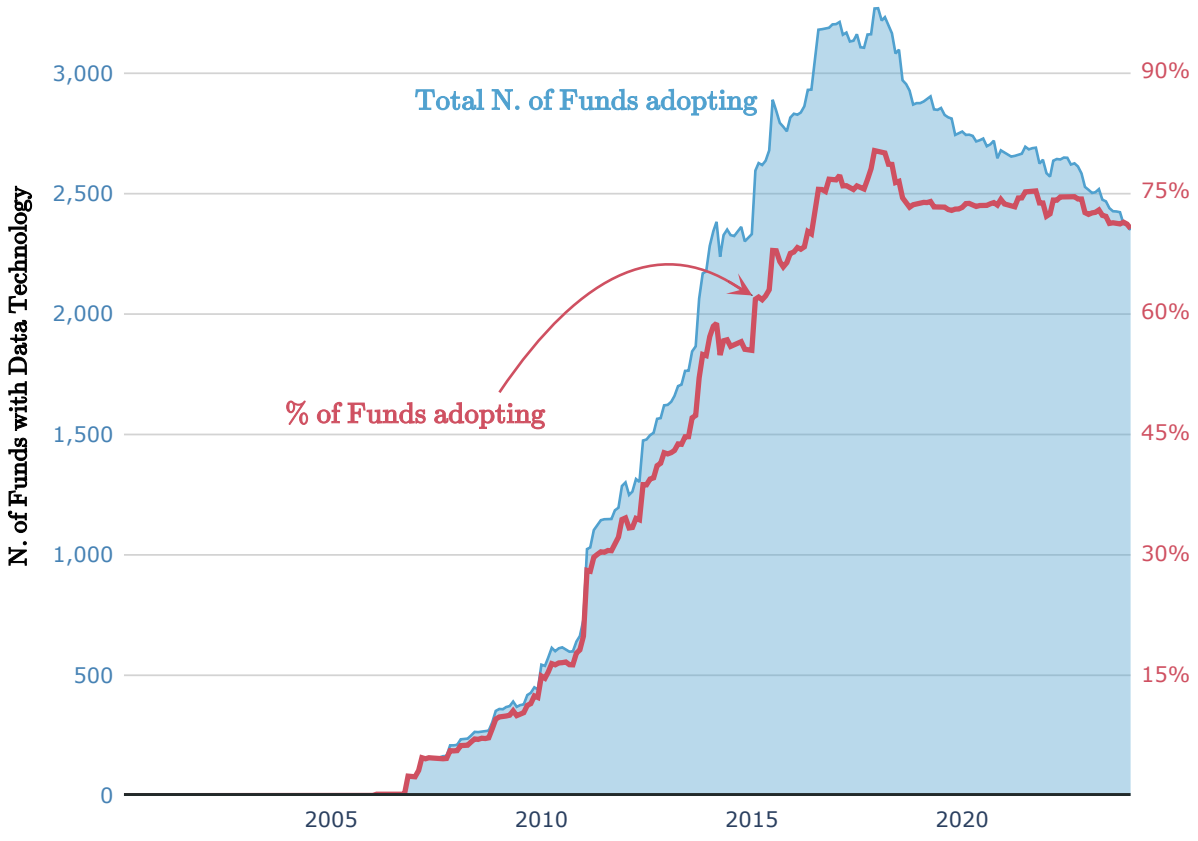


FIGURE 1: **Funds and Data Technology adoption**. This figure shows the adoption of data technologies aimed to capture visitors data, on funds' websites. The data are from BuiltWith, which detects the installation and removal of various technologies by analyzing webpage code. See Section 3.2 for details on data technologies. The blue line (left axis) represents the number of funds with at least one data technology in place for each month of the sample period. The red line (right axis) shows the percentage of funds adopting data technologies relative to the total number of funds in a given month.

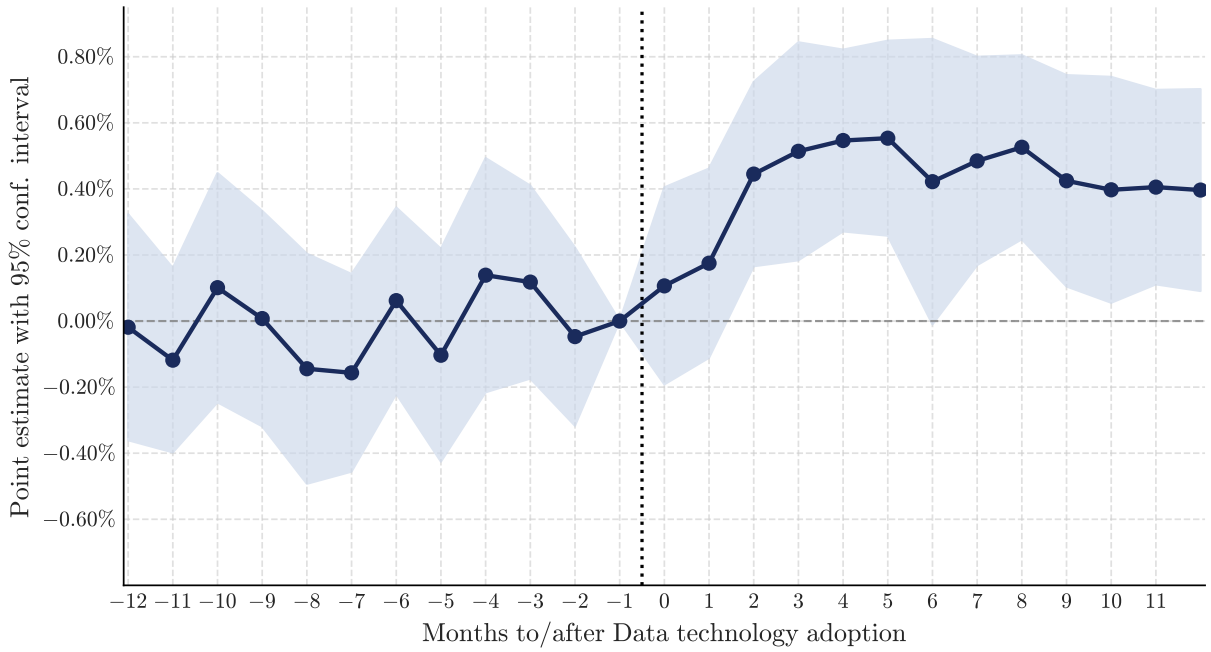


FIGURE 2: **The Dynamic Effect of Data Technologies on Fund Flows**. This figure shows results for the stacked difference-in-differences regression where the dependent variable is the one-month-ahead fund flow. Each point represents the estimated coefficient on the treatment group interaction with each month before/after data technology adoption. The treatment is a dummy equal to one if a fund i has a data technology in place at month t ($\text{DATA}_{i,t}$). The fund-month control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and alpha with respect to Vanguard index funds (Berk and van Binsbergen, 2015) in month t. Regression include fund and category×month fixed effects, and the gray area represent the 95% confidence interval for the coefficient estimates. The month just before data technology adoption (-1) is the excluded category in the regression, and is reported as zero in the figure. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023.

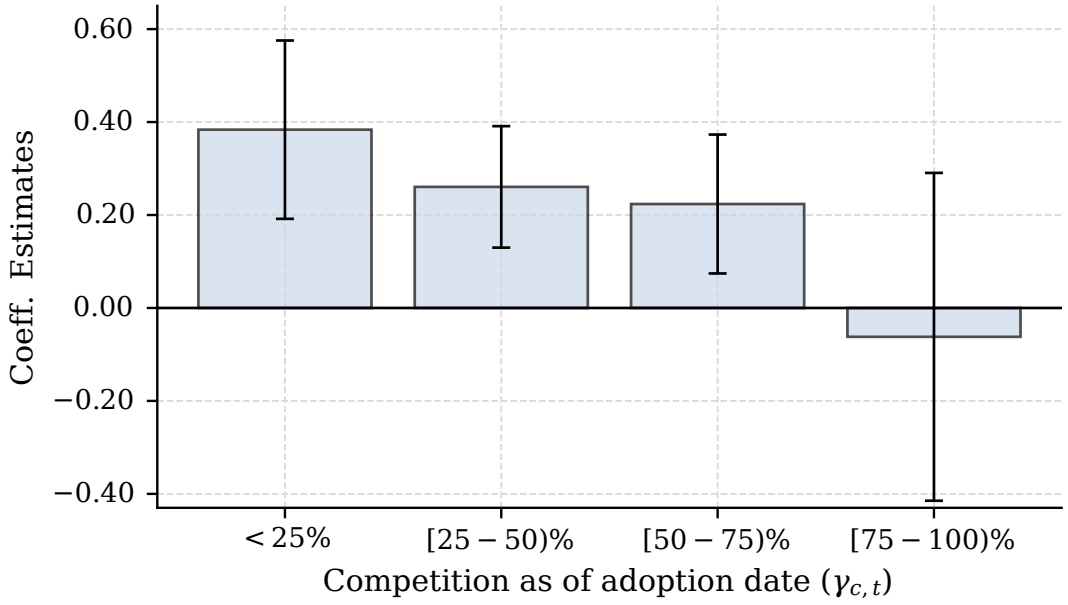


FIGURE 3: **Effect of Competition within Fund-Category**. This figure shows results for difference-in-differences coefficients across different values of competition ($\gamma_{c,t}$) as of data technology adoption. The competition coefficient $\gamma_{c,t}$ is built following equation (6), and it captures the fraction of funds with data technologies in place within fund category-month. Each bar represents a level of competition as of adoption date (e.g., the first bar represents all fund managers installing their first data technology when less then 25% of funds within its own fund-category have a data technology already installed). Each vertical line represents the 95% confidence interval. The specification is the same as the main specification in equation (2). All regressions include fund and time fixed effects, and controls: fund's size (logTNA), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t.

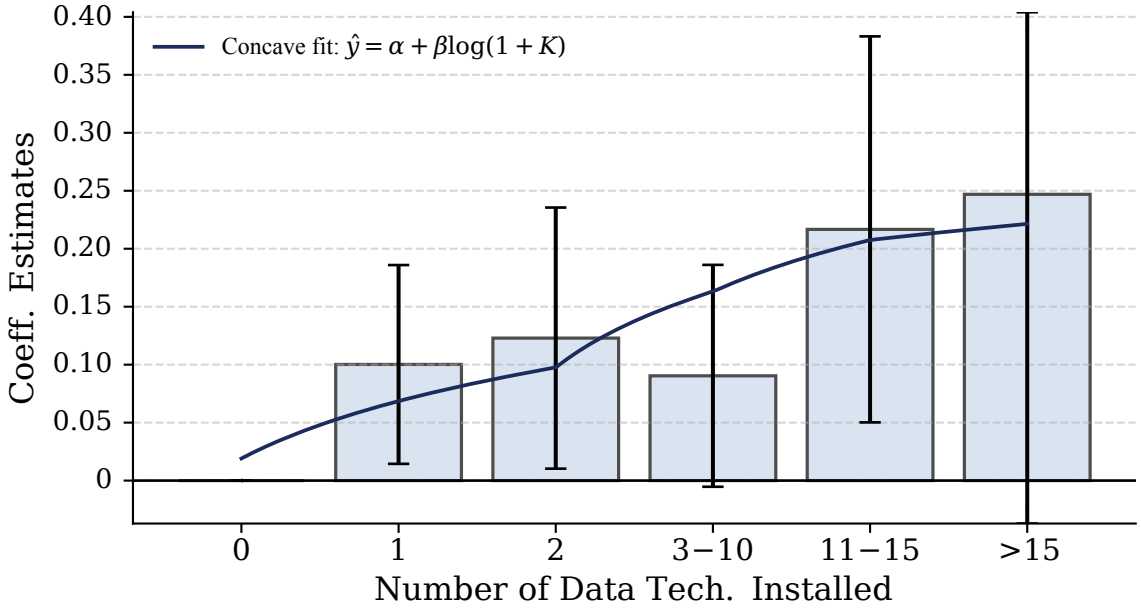


FIGURE 4: **Decreasing Marginal Benefits from Data**. This figure shows results for difference-in-differences coefficients across different bins for the number of data technologies installed. Each bar represents a bin in which fund i has K data technologies as of month t, with $K = \{0; 1; 2; [3, 10]; [11, 15]; > 15\}$. Each vertical line represents the 95% confidence interval. The solid line represents a concave fit, estimated with an OLS regression $y = \alpha + \beta log(1 + K)$, The specification is the same as the main specification in equation (2). All regressions include fund and time fixed effects, and controls: fund's size (logTNA), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t.

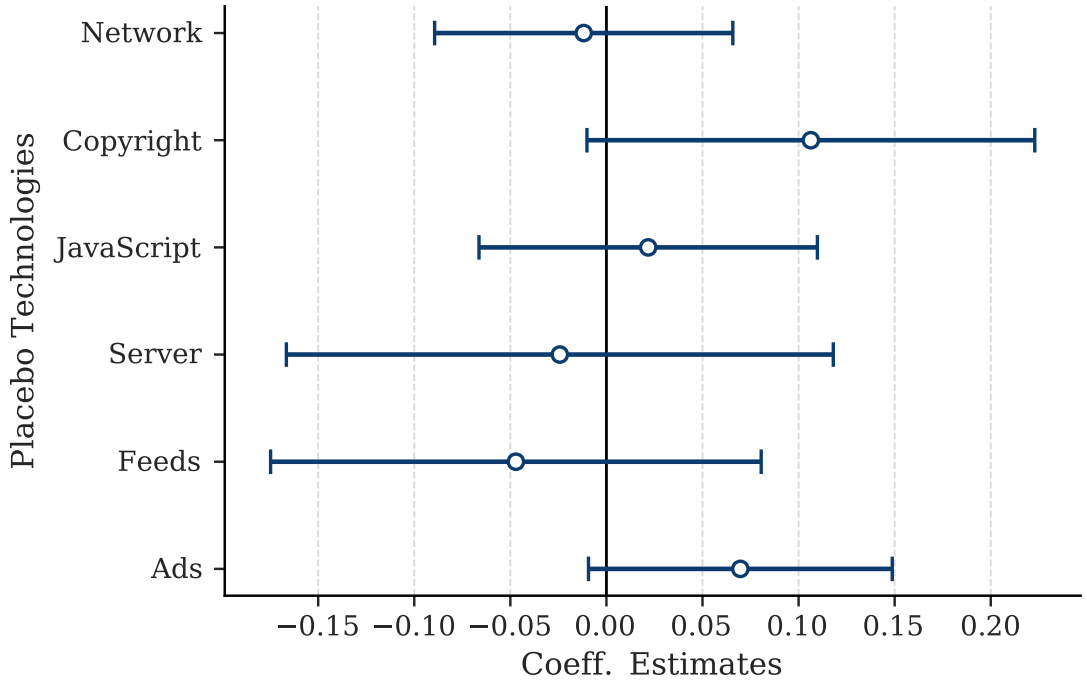


FIGURE 5: **Placebo tests**. This figure shows results from placebo tests on technologies different from data technologies. Each horizontal line represents the 95% confidence interval for tests replacing data technologies with one of the following (placebo) technologies: Network (Content Delivery Network), Server, JavaScript, Copyright, Feeds, and Ads. The specification is the same as the main results in Table 3. The dependent variable is the one-month-ahead fund flow. The confidence interval refers to the coefficient on a dummy equal to one if fund i has a placebo technology of the respective type in place at month t; i.e., analogous to θ in Equation (2). All regressions include fund and time fixed effects, and controls: fund's size (logAUM), (log) age, turnover, and CAPM alpha in month t.

	obs.	mean	sd	p5	p25	p50	p75	p95
AUM ($M)	947,079	1,177.59	2,882.62	8.69	54.93	218.65	831.99	5,821.95
Expense Ratio (%)	947,079	1.12	0.54	0.17	0.81	1.11	1.45	2.06
12b-1 Fees (%)	947,079	0.28	0.24	0.00	0.05	0.25	0.40	0.75
Flows (%)	947,079	-0.15	5.64	-5.96	-1.80	-0.65	0.78	6.98
Turnover Ratio	947,079	0.79	1.01	0.06	0.25	0.51	0.95	2.31
Age (Years)	947,079	13.15	8.84	2.58	6.00	11.08	18.75	30.17
Raw Returns	947,079	0.01	0.05	-0.08	-0.02	0.01	0.04	0.08
CAPM Alpha	947,079	-0.02	0.14	-0.30	-0.08	-0.01	0.05	0.21
$N.$ of $Data\ Tech.$	947,079	1.26	2.52	0.00	0.00	0.00	1.00	7.00
Data	947,079	0.36	0.48	0.00	0.00	0.00	1.00	1.00

Table 1: **Summary Statistics**: This table reports summary statistics for the full sample. For each variable, the table shows the number of available observations (obs,), the mean ($mean$), the standard deviation (sd), the 5th ($p5$), 25th ($p25$), 50th ($p50$), 75th ($p75$), and the 95th ($p95$) percentiles. AUM is inflation adjusted in January 2000 $ million. Expense Ratio, 12b-1 Fees, and Flows are in %; e.g., the average fund flow in the sample is -0.12% monthly. The variable $N.$ of $Data\ Tech.$ represents the total number of data technologies installed on the fund's website in a given month, the variable Data is a dummy equal to 1 if the fund-month observation has at least one data technology installed. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023.

Data Technology Name	Installation % (in 2023)	Description
Google Analytics	58.25	Users Tracking and Analytics
LinkedIn Insights	38.00	Social Media Tracking and Analytics
Adobe Analytics	37.20	Users Tracking and Analytics
Facebook Pixel	27.88	Social Media Tracking and Analytics
Omniture Test & Target	24.47	A/B Testing
RapLeaf	23.06	Users Tracking
LiveRamp	21.62	Data Connectivity Platform
Twitter Analytics	17.25	Social Media Tracking and Analytics
Bing Universal Event Tracking	16.89	Users Tracking and Analytics
mPulse	13.89	Real Time Customer Experience
Yahoo Web Analytics	12.58	Users Tracking and Analytics
Google Optimize 360	7.01	A/B Testing
Crazy Egg	6.94	Track and Visualize User Interaction
iPerceptions	6.31	Analyze Customer Feedback
Hotjar	6.05	Users Tracking and Analytics

TABLE 2: **Main Data Technologies**: This table reports the main data technologies installed on funds' websites, as of December 2023. This technologies are allow to collect and process website visitors' data. The second column shows the percentage of funds having the technology installed on its website with respect to the total number of funds, as of December 2023. The third column reports a short description of the technology's features.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.116***	0.101**	0.142***	0.130***
	(0.044)	(0.043)	(0.049)	(0.044)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266
Obs.	947,079	946,733	890,802	873,141
Adj. R^2	0.094	0.126	0.094	0.126

TABLE 3: **Fund Flows and Data Technologies**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Inflows$_{i,t+1}$ (%)		Outflows$_{i,t+1}$ (%)	
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.457***	0.419***	0.240	0.120
	(0.052)	(0.099)	(0.152)	(0.104)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	5.642	5.642	5.343	5.343
Outcome SE	15.279	15.279	15.972	15.972
Obs.	155,274	152,095	155,275	152,096
Adj. R^2	0.673	0.695	0.637	0.646

TABLE 4: **Fund Inflows, Outflows and Data Technologies**: This table shows results of panel regression on fund inflows and outflows separately, robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). I estimate equation (2) substituting the LHS with fund inflows and outflows separately. The dependent variable is the one-month-ahead fund inflows in columns (1) and (2), and fund outflows in columns (3) and (4). All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.017***	0.016***	0.037***	0.034***
	(0.003)	(0.003)	(0.005)	(0.005)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.133	1.133	1.133	1.133
Outcome SE	0.540	0.540	0.540	0.540
Obs.	947,079	946,510	890,802	873,141
Adj. R^2	0.922	0.926	0.922	0.926

Table 5: **Expense Ratio and Data Technologies**: This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,c,y+1}$ (%)		Expense Ratio$_{i,c,y+1}$ (%)		First-stage DATA$_{i,c,y}$	
	(1)	(2)	(3)	(4)	(5)	(6)
DATA$_{i,c,y}$	1.490**	1.792**	0.031***	0.032***		
	(0.732)	(0.833)	(0.009)	(0.010)		
DATA ANALYTICS GRAD.$_{c,y}$					0.011***	0.011***
					(0.003)	(0.003)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	-1.723	-1.723	1.084	1.084	0.430	0.430
Outcome SE	17.397	17.397	0.527	0.527	0.495	0.495
Obs.	40,485	39,531	40,485	39,531	40,485	39,531
Adj. R^2	0.326	0.313	0.659	0.359	0.633	0.638
F-Stat	22.522	29.268	22.522	29.268		

TABLE 6: **IV Estimates: Local Supply of Data Analytics Graduates**: This table shows results for the instrumental variable estimates where the instrument is the local supply of graduates in data analytics-related fields. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y, with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix C for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. Columns (5) and (6) show results for the first stage. Columns (1)-(2) and columns (3)-(4) report results for annual fund flows and fees, respectively. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y (DATA$_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in year y. All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)					
z_i :			Tenure of Adoption		N. of Data Tech.	
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Data}_{i,t}$	0.590***	0.606***	0.619***	0.672***	0.623***	0.643***
	(0.126)	(0.126)	(0.143)	(0.146)	(0.138)	(0.139)
$\text{Data}_{i,t} \times Post_t$	0.260**	0.313***	0.030	0.001	0.062	0.119
	(0.109)	(0.109)	(0.392)	(0.376)	(0.143)	(0.136)
$\text{Data}_{i,t} \times Post_t \times z_i$			0.051*	0.077**	0.109***	0.113***
			(0.029)	(0.032)	(0.042)	(0.039)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	0.141	0.141	0.141	0.141	0.141	0.141
Outcome SE	6.308	6.308	6.308	6.308	6.308	6.308
Obs.	770,276	769,423	584,488	583,837	689,466	688,820
Adj. R^2	0.107	0.139	0.094	0.132	0.101	0.133

TABLE 7: **Fund Flows and Data Technologies after TensorFlow Release**: This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead fund flow. Columns (1) and (2) follow specification in equation (3), while columns (3) to (6) follow (4). In columns (3) and (4) the continuous treatment z_i is the (log) number of months between the first data technology adoption and TensorFlow's release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow's release, as continuous treatment z_i. $\text{Data}_{i,t}$ is a dummy equal to one if fund i has a data technology in place at month t. See Section 3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (6)) in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow's release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Share Class Flows$_{j,i,t+1}$ (%)			
	(1)	(2)	Only Retail (3)	Only Institutional (4)
Data$_{i,t}$	0.059	0.047	0.120**	-0.019
	(0.085)	(0.084)	(0.052)	(0.097)
Data$_{i,t}$ × Retail	0.200***	0.175**		
	(0.071)	(0.069)		
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	✓
Category×Time FE	×	✓	×	×
Outcome mean	0.631	0.631	0.383	0.691
Outcome SE	10.502	10.502	7.896	8.353
Obs.	799,667	798,719	334,280	150,731
Adj. R^2	0.082	0.092	0.123	0.091

Table 8: **Retail and Institutional Share Classes**: This table shows results of panel regression in which the dependent variable is the one-month-ahead flow for share class j of fund i. The regressors are a dummy equal to one if fund i has a data technology in place at month t (Data$_{i,t}$) interacted with the share class' j type (retail or institutional), 12b-1 fees, and controls for share class-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. Columns (3) use only retail share classes observations, while columns (4) only institutional share classes observations. The control variables include a share class' (log) AUM, (log) age, flows, turnover, and CAPM alpha in month t. The monthly sample is from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$w_{i,q}$(cash) (%)				Amihud$_{i,q+1}$ (%)		Redemption Fee	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$D/P_{i,q}$	-0.379***	-0.289*	-0.226	-0.123				
	(0.131)	(0.172)	(0.139)	(0.185)				
Data$_{i,q} \times D/P_{i,q}$			-0.585**	-0.648**				
			(0.236)	(0.324)				
Data$_{i,q}$			0.001	0.001	0.405***	0.362***	-0.017***	-0.016*
			(0.001)	(0.001)	(0.007)	(0.094)	(0.006)	(0.009)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	4.355	4.355	4.355	4.355	0.448	0.448	0.240	0.240
Outcome SE	12.226	12.226	12.226	12.226	3.946	3.946	0.427	0.427
Obs.	209,868	209,720	209,868	209,720	158,027	154,927	254,359	251,328
Adj. R^2	0.552	0.555	0.552	0.555	0.310	0.330	0.745	0.756

TABLE 9: **Liquidity Management and Data Technology**: This table shows results of panel regression in which the dependent variable is the quarterly fund's cash holdings (columns (1) to (4)), the portfolio Amihid illiquidity ratio (columns (5)-(6)), and the likelihood that the fund charges redemption fees to investors (columns (7)-(8)). In columns (1) to (4) the regressors are the fund's portfolio dividend-price ratio ($D/P_{i,q}$), an interaction term with a dummy equal to one if a fund i has a data technology in place in quarter q (Data$_{f,t}$), and controls for fund-quarter characteristics. See Section 3.2 for details on data technologies. The fund-quarter control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, and 12b-1 fees. Portfolio holdings data are from Thomson Reuters (s12). The quarterly sample is from 2004Q2 to 2023Q4. All standard errors are two-way clustered by fund and quarter (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)				Expense Ratio$_{i,t+1}$ (%)			
	Active		Passive		Active		Passive	
	(1)	(2)	(3)	(4)	(5)	(6)	(6)	(8)
Data$_{i,t}$	0.138***	0.113**	0.578***	1.160**	0.034***	0.035***	0.024**	0.044**
	(0.053)	(0.051)	(0.210)	(0.517)	(0.006)	(0.006)	(0.011)	(0.020)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	0.086	0.086	0.976	0.976	1.186	1.186	0.726	0.726
Outcome SE	5.745	5.745	9.311	9.311	0.504	0.504	0.633	0.633
Obs.	726,304	712,701	84,571	71,759	726,304	712,701	84,572	71,760
Adj. R^2	0.111	0.148	0.056	0.081	0.899	0.904	0.973	0.975

Table 10: **Active and Passive Funds**: This table shows results of panel regression in which the dependent variable is the one-month-ahead flow (columns (1) to (4)) or expense ratio (columns (5) to (8)). In columns (1), (2), (5), and (6) I only include active funds, while columns (3), (4), (7), and (8) only passive funds —including ETFs. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y (Data$_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in year y. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	12b-1 Captive Sales$_{i,t+1}$		Marketing-Sales Exp.$_{i,t+1}$	
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.001***	0.001***	0.007***	0.008***
	(0.000)	(0.000)	(0.002)	(0.001)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.003	0.003	0.119	0.119
Outcome SE	0.015	0.015	0.142	0.142
Obs.	180,723	178,046	252,207	249,139
Adj. R^2	0.760	0.761	0.734	0.738

Table 11: **Captive Retail Sales Force**: This table shows results of panel regression on funds' captive retail sales forces, robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). The dependent variable is the fund's one-month-ahead payment to captive retail sales force from 12b-1 fees, over its AUM (columns (1)-(2)). While in columns (3)-(4) the dependent variable is the fund's total marketing and sales costs over its AUM. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
Marketing-Sales Exp.$_{i,t}$	0.188***	0.152**	0.103	0.061
	(0.062)	(0.058)	(0.065)	(0.061)
Marketing-Sales Exp.$_{i,t}$× Data$_{i,t}$			0.223***	0.234***
			(0.053)	(0.053)
Data$_{i,t}$			2.505***	2.523***
			(0.595)	(0.591)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.266	0.266	0.266	0.266
Outcome SE	6.529	6.529	6.529	6.529
Obs.	35,358	34,507	35,358	34,507
Adj. R^2	0.195	0.230	0.197	0.232

Table 12: **Marketing Expenditures and Fund Flows**: This table shows results of panel regression on funds' marketing and sales efforts after adoption of a data analytics technology. The dependent variable is the fund's cumulative six-months-ahead fund flows. The regressors are a fund's (log) total marketing and sales expenditures, a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), the interaction between Data$_{i,t}$ and (log) marketing expenditures, and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. The semi-annual sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and calendar semester (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	FRE Index		deHaan et al. (2021)		Second-person		Mention Themes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\text{DATA}_{i,y}$	0.841***	0.781***	-1.042***	-1.063***	0.018***	0.015***	0.021*	0.016**
	(0.219)	(0.156)	(0.077)	(0.142)	(0.002)	(0.004)	(0.011)	(0.008)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	22.093	22.093	27.508	27.508	0.051	0.051	0.259	0.259
Outcome SE	93.263	93.263	88.601	88.601	0.223	0.223	0.629	0.629
Obs.	111,000	98,345	101,695	98,347	101,694	98,343	111,004	108,374
Adj. R^2	0.155	0.155	0.152	0.149	0.353	0.357	0.417	0.417

TABLE 13: **Text Analysis on Fund Prospectuses**: This table shows results of panel regression in which the dependent variables are from textual analysis of fund prospectuses. Columns (1)-(2) use the Flesch Reading Ease (FRE) index (Flesch, 1948), columns (3)-(4) report results for the average words per sentence in the prospectus deHaan et al. (2021), columns (5)-(6) for the frequency of second-person pronouns (e.g., "you", "your"), and columns (7)-(8) for the occurrency of words related to themes (Ben-David et al., 2023). I refer to Appendix D for details on the construction of text-analysis measures. The fund-year control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, and 12b-1 fees. Fund prospectuses are from SEC's EDGAR (see Appendix B). The annual sample is from 2006 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$\bar{d}$		iVol		Active Share	
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{DATA}_{i,q}$	0.004*	0.001	-0.001*	-0.001*	-0.004*	-0.003
	(0.002)	(0.003)	(0.000)	(0.000)	(0.002)	(0.002)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	-1.723	-1.723	1.084	1.084	0.430	0.430
Outcome SE	17.397	17.397	0.527	0.527	0.495	0.495
Obs.	158,133	155,034	158,133	155,034	158,132	155,033
Adj. R^2	0.784	0.817	0.789	0.829	0.602	0.621

TABLE 14: **Product Differentiation for Existing Funds**: This table shows results of panel regression in which the dependent variables are measures of funds product differentiation based on portfolio holdings. Columns (1)-(2) use the holdings distance measure of Hoberg et al. (2017), columns (3)-(4) report results for the portfolio idiosyncratic volatility, and columns (5)-(6) for active share (Cremers and Petajisto, 2009). The fund-quarter control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, and 12b-1 fees. Holdings data are from Thomson Reuters (s12) (see Appendix B). The quarterly sample is from 2004Q2 to 2023Q4. All standard errors are two-way clustered by fund and quarter (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	log(N. of Funds)$_{f,t+1}$		Thematic words in PIS:	All		AI & Crypto		ESG	
	(1)	(2)		(3)	(4)	(5)	(6)	(7)	(8)
Data$_{f,t}$	0.020**	0.034**		0.104***	0.071***	0.005***	0.002	0.085***	0.055***
	(0.009)	(0.017)		(0.012)	(0.013)	(0.001)	(0.002)	(0.011)	(0.012)
Fund Family FE	✓	✓		×	×	×	×	×	×
Time FE	✓	✓		×	✓	×	✓	×	✓
Outcome mean	1.485	1.485		0.157	0.157	0.004	0.004	0.115	0.115
Outcome SE	1.350	1.350		0.438	0.438	0.055	0.055	0.402	0.402
Obs.	124,560	124,560		5,568	5,568	5,568	5,568	5,568	5,568
Adj. R^2	0.885	0.914		0.013	0.047	0.002	0.009	0.011	0.045

Table 15: **Launch of New Funds**: This table shows results of panel regression in which the dependent variable is the number of funds offered by fund family f in month $t+1$ in columns (1)-(2). The regressors are a dummy equal to one if at least one fund within family f has a data technology in place at month t (Data$_{f,t}$), and the (log) fund family age. See Section 3.2 for details on data technologies. In columns (3) to (8), I only use observations on the first prospectus released by *all* newly launched funds in the sample. Columns (3) to (8) regress mentions of words related to retail-oriented themes such as AI, ESG, Cannabis, cybersecurity, political values, space, and video games in the newly launched funds' Principal Investment Strategy (PIS) of their prospectuses. I detail all words I use to identify these themes in Appendix D. Columns (3)-(4) consider all thematic words. Columns (5)-(6) consider only AI and Crypto-related words, while columns (7)-(8) uses mentions of ESG words in new funds' prospectuses. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Appendix

A Economic Framework

A.1 Equilibrium and Main Predictions

To illustrate the main mechanism, I develop a tractable framework of fund flows and fees in a competitive, rational market. The setup builds on Berk and van Binsbergen (2015). I introduce the assumption that managers compete not only on performance, but also on how well they understand customer preferences. I then derive the equilibrium implications for funds' assets under management (AUM) and the split of rents (from data) between asset managers and investors.

The model features a continuum of investors who allocate their wealth to a measure-one continuum of asset managers, which I index by i. Time is discrete and infinite. Each manager i at time t has AUM $q_{i,t}$, and charges a per-dollar fee $f_{i,t}$. Fund managers are Bayesian.

I assume investors have (homogenous) non-pecuniary preferences θ_t that follow an AR(1) process:

$$\theta_{t+1} = \rho\theta_t + \varepsilon_{t+1}, \qquad \rho \leq 1$$

with $\varepsilon_{t+1} \sim \mathcal{N}\left(0, \tau_\theta^{-1}\right)$. θ_t may capture distribution channel preferences, such as whether an investor prefer to buy product directly vs. trough financial adviser, or communication frequency (daily vs. monthly communication). It may also capture a non-pecuniary preference for assets with particular characteristics. For instance, some investors might have non-pecuniary preference to hold green stocks (Hartzmark and Sussman, 2019), politically exposed stocks, or specific themes (Ben-David et al., 2023). Crucially, asset managers do not observe θ_t directly. Instead, each manager i receives a noisy unbiased signal $s_{j,t} = \theta_t + \eta_{j,t}$, with $\eta_{j,t} \sim \mathcal{N}\left(0, \tau_s^{-1}\right)$. Where $\eta_{j,t}$ are i.i.d and independent of ε_t. The signal precision is τ_s.

A fraction $\lambda \in (0, 1]$ of asset managers, called "data-managers", observe an additional independent signal from analyzing customers' data: $s_{D,t} = \theta_t + \eta_{D,t}$, with $\eta_{D,t} \sim \mathcal{N}\left(0, \tau_D^{-1}\right)$. D stays for data-managers. Data-managers thus combine two independent signals and have total precision $\tau_s + \tau_D$. I treat λ as given here, and endogenize it in Section A.3.

Managers update beliefs about θ_t using standard recursive Bayesian updating (Kalman filter):

$$\begin{aligned}\hat{\theta}_{j,t+1} &= \rho\hat{\theta}_{j,t} + K_{j,t}\left(s_{j,t} - \hat{\theta}_{j,t}\right)\\ K_{j,t} &= \rho\frac{\tau_j}{\hat{\Sigma}_{j,t}^{-1} + \tau_j}\\ \hat{\Sigma}_{j,t+1} &= \rho^2\frac{1}{\hat{\Sigma}_{j,t}^{-1} + \tau_j} + \frac{1}{\tau_\theta},\end{aligned} \tag{A.1}$$

where $\tau_j = \tau_s + \mathbb{1}_{\{j\in D\}}\tau_D$. I denote $\hat{\theta}_t = \mathbb{E}[\theta_t|s_0,\ldots,s_{t-1}]$ and $\hat{\Sigma}_t = \mathbb{V}ar[\theta_t|s_0,\ldots,s_{t-1}] = \mathbb{E}[(\theta_t - \hat{\theta}_t)^2]$.

Following the data economy literature (e.g., Farboodi and Veldkamp, 2020; Abis and Veldkamp, 2023; Cong et al., 2021), I define a manager's *stock of knowledge* as the conditional precision:

$$\Omega_{i,t} = \hat{\Sigma}_{j,t}^{-1} = \mathbb{E}[(\theta_t - \hat{\theta}_t)^{-2}]. \tag{A.2}$$

This stock measure summarizes the total knowledge a manager accumulates about θ_t by observing $s_{j,0}, s_{j,1}, \ldots, s_{j,t}$.

Investors derive per-period utility from investing in fund i at time $t-1$:

$$\text{TP}_{j,t} = a - b\cdot q_{i,t} - f_{i,t} + \kappa\cdot(\theta_t - \hat{\theta}_{j,t})^{-2}. \tag{A.3}$$

The first term, $a - b\cdot q - f$, is the same as in Berk and van Binsbergen (2015). Managers generate gross alpha with decreasing returns to scale ($a - b\cdot q$) and charge a fee (f) to investors holding the fund. The second term is new. Investors gain non-pecuniary utility from holding funds closer to their preference. The parameter κ governs the importance of this non-pecuniary component. For example, ESG consumers might derive non-pecuniary utility from buying green-labeled products. If an asset manager recognizes this taste, she can cater and attract those investors. This second term is what makes data relevant in the model.

Standard neoclassical assumptions hold: investors are rational, markets are competitive, and managers maximize profits (Berk and Green, 2004; Berk and van Binsbergen, 2015). In equilibrium, investors supply capital to funds and the expected utility from each fund must

equal zero:

$$\mathbb{E}[\text{TP}_{j,t}] = 0, \quad \forall\, j. \tag{A.4}$$

In this rational competitive market for asset management services, the equilibrium AUM is proportional to the stock of knowledge about investors' preference:

$$q^*_{j,t} = \frac{a + \kappa\Omega_{i,t}}{2b}. \tag{A.5}$$

Asset managers with more data (larger Ω) attract more AUM. Therefore, managers that know more about investors' non-pecuniary preferences are larger, all else equal. Equilibrium fund flows follows directly:

$$Flow_{j,t} = \frac{\kappa}{2b}(\Omega_{i,t+1} - \Omega_{i,t}) = \frac{\kappa}{2b}\Delta\Omega_{i,t+1}. \tag{A.6}$$

Thus, managers with more (absolute) improvement in their stock of knowledge $\Delta\Omega_{t+1}$ receive more flows. Data-managers, who crunch customers' data, attract more flows relative to traditional managers. This yields hypothesis 1 in the main text.[37]

Proof. (*Hypothesis 1*)

The equilibrium condition (A.4) writes:

$$\begin{aligned} a - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot \mathbb{E}[(\theta_t - \hat{\theta}_{j,t})^{-2}] &= 0 \\ a - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot \Omega_{i,t} &= 0 \\ \rightarrow q_{i,t}(f_{i,t}) &= \frac{a - f_{i,t} + \kappa \cdot \Omega_{i,t}}{b} \end{aligned} \tag{A.7}$$

Where in the second row I used $\Omega_{i,t} = \hat{\Sigma}^{-1}_{j,t} = \mathbb{E}[(\theta_t - \hat{\theta}_t)^{-2}]$. The AUM of a fund is a linear function of the fee it charges. Let now define fund i profits at time t as $\pi_{j,t} := q_{i,t} \cdot f_{i,t}$. As

[37] Equation (A.6) also implies that stronger decreasing returns to scale (DRS) reduce the flow benefit of data. Appendix Figure E.4 confirms this prediction. I follow the recursive demeaning approach in Pástor, Stambaugh and Taylor (2015) to estimate fund-specific DRS coefficients as-of adoption of a data technology. Then, I plot the effect of data technologies on flows for different bins of DRS. As predicted by the theoretical framework, the relationship between the effect of data on flows, and DRS is monotonic and decreasing. By contrast, equation (A.8) implies no relationship between data technology and fees. I show this is indeed the case in Appendix Figure E.5.

funds a maximize profit:

$$\begin{aligned} &\max \ \pi_{j,t} \\ \text{FOC:} \qquad 0 &= \frac{a - 2f^*_{j,t} + \kappa \cdot \Omega_{i,t}}{b} \\ f^*_{j,t} &= \frac{a + \kappa \cdot \Omega_{i,t}}{2} \end{aligned} \tag{A.8}$$

This optimum is unique and global. The second order condition $-2/b < 0$ is always satisfied. Finally, substituting (A.8) in (A.7) yields the equilibrium AUM (A.5), from which (A.6) follows directly. □

Hypothesis 2 follows directly from equation (A.7). Profit-maximizing managers charge higher fees, as they better what customers prefer. This result is akin to firms charging higher prices for products that elicit higher willingness to pay. Selling specialized products allows firms to command higher prices (Menzio, 2023), a common result in IO several models (Salop, 1979; Lancaster, 1966; Pellegrino, 2024).

Difference-in-Differences Coefficients. In the main text, I test these hypotheses using a difference-in-differences specification. The difference-in-differences coefficients can be interpreted through the lens of this theoretical framework.

Consider two identical funds, A and B, with the same AUM, fees, and knowledge at time t, so $\Omega_{A,t} = \Omega_{B,t} = \Omega_t$. Assume that at time t^+, fund A receives exogenous access to more precise signals, i.e., it A becomes a data-manager. Manager A now observes two independent signals with total precision $\tau_s + \tau_D$, while manager B receive one signal whose precision is τ_s. In equilibrium, the two funds receive flows:

$$\begin{aligned} Flow_{A,t} &= \frac{\kappa}{2b}(\Omega_{A,t+1} - \Omega_t) \\ Flow_{B,t} &= \frac{\kappa}{2b}(\Omega_{B,t+1} - \Omega_t). \end{aligned}$$

The difference-in-differences coefficient on flows, estimated in the sample $t \in [0, t+T]$ is:

$$\beta^{DiD}_{Flows,t\to t+T} = \frac{\kappa}{2b}(\Omega_{A,t+T} - \Omega_{B,t+T}). \tag{A.9}$$

Therefore, the more precise the additional signal A receives, the larger is β^{DiD}_{Flows}. Moreover,

equation (A.9) predicts that the diff-in-diff coefficient on flows is smaller for funds facing larger decreasing returns to scale —i.e., funds for which b is large. Appendix Figure E.4 confirms that this is the case: β^{DiD}_{Flows} decreases monotonically with the decreasing returns to scale parameter b.

Similarly, the difference-in-differences coefficient on fees is:

$$\beta^{DiD}_{Fees,t\to t+T} = \frac{\kappa}{2}(\Omega_{A,t+T} - \Omega_{B,t+T}). \tag{A.10}$$

Again, a more precise signal from data analytics technologies is associated with a higher β^{DiD}_{Fees}. Unlike flows, however, β^{DiD}_{Fees} does not depend on decreasing returns to scale. Appendix Figure E.5 confirms that the effect of data adoption on fees is flat across funds facing different decreasing returns to scale.

A.2 Value Added

From this theoretical framework, I can follow Berk and van Binsbergen (2015) in defining the total value added that a fund manager generates as:

$$VA_{j,t} = q^*_{j,t} \cdot f^*_{j,t} = \frac{(a + \kappa \cdot \Omega_{i,t})^2}{4b}. \tag{A.11}$$

Funds with more knowledge about customers' preference generate higher value added than otherwise equivalent funds. Importantly, this additional value does not come from value that funds extract from financial markets. Instead, it arises from better matching the product they offer to investors' non-pecuniary preferences.

To show this, I decompose total value added into two components. The first is the value extracted from financial markets, VA^{FIN}, equivalent to equation (5) in Berk and van Binsbergen (2015):

$$VA^{FIN}_{j,t} = q^*_{j,t} \cdot \alpha^G_{jt,} = \frac{a^2 - \kappa^2 \cdot \Omega^2_{i,t}}{4b}. \tag{A.12}$$

The second component is new: the value added from analyzing customers' data, VA^{DATA}. This term captures the value fund managers create by tailoring products to investor prefer-

ences:

$$VA_{j,t}^{DATA} = \frac{a\kappa \cdot \Omega_{i,t} + \kappa^2 \cdot \Omega_{j,t}^2}{2b}. \tag{A.13}$$

As expected, value added from data increases with a fund's stock of knowledge. Asset managers who collect and process customers' data can better align funds with customers' taste, generating higher utility for their investors, via a better product match.

The *total* value added rises with stock of knowledge, but its composition shifts. The share of value added extracted from financial markets declines with funds' accumulating more data. While the share coming from analyzing data grows. As funds rely more and more on customers' data, data-driven value added crowds out the value managers generates investing in financial markets. Figure A.1 illustrates this decomposition as the stock of knowledge grows.

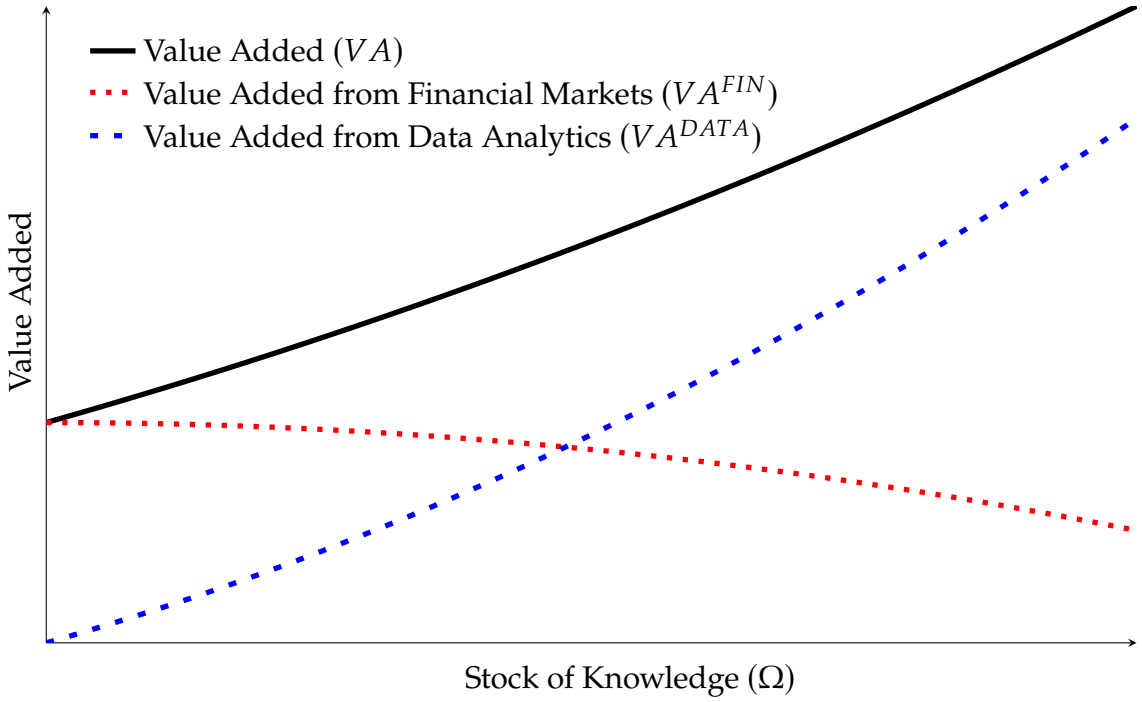


FIGURE A.1: **Value Added and Stock of Knowledge**. The figure shows the value added a fund manager generates, as its stock of knowledge (data) grows. The black solid line depicts the total value added generated by an asset manager. The red dotted line shows the value added by extracting value from financial markets, while the blue dashed line represents the value added by analyzing data.

A fund manager's incentive to deliver performance declines as its stock of knowledge grows. Funds' AUM increases and decreasing returns to scale erode the risk-adjusted

performance they deliver to investors. Yet, equilibrium AUM continues to rise because investors allocate more wealth when fund offerings align with their preferences, even absent alpha. This effect shifts the incentives for fund managers from generating alpha, to selling products that match investors' taste. This echoes Ben-David et al. (2023), who show that thematic ETFs deliver negative risk-adjusted performance after launch, on average.

A.3 Endogenous Choice of Adopting Data Technologies

I next endogenize managers' choice to adopt a data technology. Before each period t, a manager can pay a per-period cost $\bar{c}$ to acquire a data analytics technology.

Paying this cost grants access to an additional unbiased signal $s_{D,t} = \theta_t + \eta_{D,t}$, with $\eta_{D,t} \sim \mathcal{N}\left(0, \tau_D^{-1}\right)$. Managers also differ in skill a_j, which is the alpha a manager delivers on the first dollar invested. Higher a_j represents greater skill. Each manager has skill drawn from $a_j \sim \mathcal{N}(\bar{a}, \tau_a^{-1})$.

Investors' utility remains:

$$\text{TP}_{j,t} = a_j - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot (\theta_t - \hat{\theta}_{j,t})^{-2}. \tag{A.14}$$

As before, asset managers filtering better investors' preference, θ_t, deliver higher utility all else equal. Managers who pay $\bar{c}$ observe more precise signals, match preferences more closely, and attract more capital. From equations (A.8) and (A.5), fund manager i's optimal profits are

$$\pi^*_{j,t} = \frac{(a_j + \kappa \cdot \Omega_{i,t})^2}{4b} := \pi_k(a_j) \quad k \in \{D, T\}, \tag{A.15}$$

where $k \in \{D, T\}$ denotes whether the fund decides to adopt data analytics and being a "data-manager" (D), or not (T). In equilibrium, the benefit from adoption must equal the cost $\bar{c}$. I first compute the gross profit gain from data analytics is:

$$\begin{aligned} \pi_D(a_j) - \bar{c} = \frac{(a_j + \kappa \cdot \Omega^D_{j,t})^2 - (a_j + \kappa \cdot \Omega^T_{j,t})^2}{4b} = \\ = \frac{\kappa^2[(\Omega^D_{j,t})^2 - (\Omega^T_{j,t})^2]}{4b} + \frac{\kappa(\Omega^D_{j,t} - \Omega^T_{j,t}) \cdot a_j}{2b} \end{aligned} \tag{A.16}$$

The important intuition here is that the (gross) benefits of adoption increase with skill a_j. More skilled managers benefits more from data analytics. However, there exist a threshold a^* below which managers do not adopt, because those without enough skills ($a_j < a^*$) earn too little profits to cover costs $\bar{c}$. This threshold is a^* such that $\pi_D(a^*) - \bar{c} = 0$. Solving for it, gives the:

$$a^* = \underbrace{\frac{2b\bar{c}}{\Omega^D_{j,t} - \Omega^T_{j,t}}}_{\text{cost/benefit}} - \frac{\Omega^D_{j,t} + \Omega^T_{j,t}}{2}, \tag{A.17}$$

and a manager i adopts a technology to analyze customers' data if $a_j \geq a^*$, whereas when $a_j < a^*$ the manager does not adopt. The first term in equation (A.17) captures cost over benefit of adopting. If adoption is expensive (high $\bar{c}$) or the informational gain is small, no managers adopt. For instance, this may happen because it is very costly to acquire human capital trained in data analytics (high $\bar{c}$), or because the informational benefit of analyzing data is too little. On the other hand, when the cost-benefit term is small, many funds will install data technologies.

The second term reflects pure rents from data. This component comes from the fact that even low-skilled managers gain profits from analyzing data. To see this, notice that when a manager has no skill ($a_j = 0$), she can still earn positive profits (from data): $\pi_D(0)\frac{(\kappa\cdot\Omega)^2}{4b}$. As a consequence, the skill threshold a^* that justify investment in data analytics is lower when this pure rent term is large. The equilibrium share of adopters is

$$\lambda = \mathbf{P}\{a_j \geq a^*\} = 1 - \Phi\left(\frac{a^* - \bar{a}}{\tau_a^{-1}}\right). \tag{A.18}$$

In Section 4.3.2, I use the local supply of data analytics graduates to instrument adoption of a data technology. The intuition is that local experts reduce cost of adoption (first term in (A.17)), as it lowers the threshold a^* and raises the share of adopters.

B Data Appendix

In this Appendix I describe the main dataset construction procedure. I use six main data sources: (i) CRSP Survivorship-Bias-Free US Mutual Funds data, (ii) FactSet Funds data, (iii) Morningstar Direct, (iv) BuiltWith for websites' technology installation/removal dates, (v) N-SAR regulatory filings, and (vi) mutual funds portfolio holdings from Thomson Reuters (s12).

B.1 CRSP Mutual Funds

I follow Berk and van Binsbergen (2015) and Pástor, Stambaugh and Taylor (2015) procedures as closely as possible. I start from the raw CRSP Survivorship-Bias-Free US Mutual Funds monthly data. Each observation in this dataset identifies a fund's share class ($crsp_fundno$) in a given month. The raw CRSP Mutual Funds dataset from January 1980 to December 2023 has 9,327,753 share class-month observations. I start filling missing contact information in CRSP data; i.e., $address1$, $city$, $state$, $website$, and zip. I fill missing contact information between two (or more) non-empty contact information within the same share class, when the two non-empty entries coincides. This step replaces 62,266 missing obs. (0.68% of total) with non-empty entries.

CRSP reports a fund's $website$ starting January 2008. I use information from `whois.com` to backward fill missing websites' observations before January 2008. In particular, `whois.com` has information on websites' registration date, hosting service, and other characteristics. I hand-collected from `whois.com` the registration date for each website in CRSP, and I verify the website belongs to the CRSP fund using the Internet Archive Wayback Machine. Then, I backward fill missing $website$ observations included between the website's registration date and January 2008. This procedure backward fill 1,423,338 obs. (15.59% of total).

Following Berk and van Binsbergen (2015), I backfill missing CUSIP with the last available non-empty CUSIP within the same share class. This step replace 871, 728 CUSIP obs. (9.55% of total). I do not *forward* fill missing observations. I replace missing exp_ratio and $actual_12b1$ fees with their time series average within the same share class (Roussanov, Ruan and Wei, 2020). This step fills 646,928 obs. (7.09% of total) and 1,510,034 obs. (16.54%

of total) respectively. Following Sirri and Tufano (1998) and Roussanov, Ruan and Wei (2020) I compute the "effective" 12b-1 fee summing CRSP's $actual_12b1$ to the share class-month front load, and assuming the front load fee is amortized over 7 years.
I adjust AUM (TNA) for inflation to be comparable across time. The seasonally adjusted monthly CPI is from FRED All Consumers: All Items, and I use January 2000 as baseline month (as in Berk and van Binsbergen, 2015).

Finally, since several funds in CRSP report their AUM only at the quarterly (or annual) frequency before March 1993 (Pástor, Stambaugh and Taylor, 2015), I drop all share class-month observations before that date. I also drop observations with missing CUSIP and ticker. After this steps, I have 8,825,188 share class-month observations from the CRSP Mutual Funds dataset.

B.2 FactSet Funds

I obtain mutual funds data from FactSet at the share class level, and I will merge it with CRSP data at the CUSIP-month level. I use CUSIP rather than ticker, because the CUSIP cannot be re-assigned. I use FactSet mainly to identify all share classes of the same mutual fund $FactSet_fund_id$ in a given month. From FactSet, I obtain the fund id, -fund type (e.g., ETF, Open-end fund, etc.), the fund name, brand, share class, leverage factor, category, minimum initial investment, and cash holdings.

I adjust AUM (TNA) anc cash holdings for inflation to be comparable across time (CPI from FRED All Consumers: All Items, and I use January 2000 as baseline month). I have 38,715,834 share class-month observations from the FactSet Funds dataset.

B.3 Morningstar Direct

I complement CRSP Mutual Funds and FactSet data with information on US-domiciliated mutual funds from Morningstar Direct. I use monthly data at the share class level ($secid$). The raw Morningstar Direct dataset has 5,003,970 share class-month observations, and I will merge it with CRSP/FactSet data at the ticker-month level. Following Berk and van

Binsbergen (2015), I start filling empty ticker observations with the last available non-empty ticker within share class. After this step, I set to "missing" all observations with more than one fund associated with the same ticker-month match, to avoid matching mistakes. Then, I drop all observations without a valid ticker (1,278,877 obs., 25.56 % of the initial raw data).

I adjust AUM (TNA) for inflation to be comparable across time. The seasonally adjusted monthly CPI is from FRED All Consumers: All Items, and I use January 2000 as baseline month (as in Berk and van Binsbergen, 2015).
After this steps, I have 3,541,168 share class-month observations from Morningstar Direct.

B.4 CRSP-Morningstar-FactSet Merged

I merge CRSP and FactSet dataset by CUSIP-month. Then, I merge the remaining share class-month observations with Morningstar by ticker-month. I prefer merging by CUSIP rather than ticker, because the CUSIP cannot be re-assigned.

The CRSP dataset resulting from Appendix Section B.1 has 2,954 observations (0.03% of total) in which the same CUSIP-month pair appears twice. Inspecting those observations, they do not appear to be double reporting, but rather mistakes on CRSP's side. For each of those observations, I keep the CUSIP-month with the largest total AUM in the sample.
I merge CRSP and FactSet dataset by CUSIP-month. The merge results in 7,646,161/8,689,175 observations merged (88%).

Then, I merge the remaining 1,043,024 observations with Morningstar by ticker-month. The merge results in 28,195/1,043,024 extra matches (2.70%). I drop the remaining 1,014,829 unmerged observations, since fuzzy attribution of CUSIP or ticker might result in incorrect mergers.

I classify index funds following Berk and van Binsbergen (2015) and Pástor, Stambaugh and Taylor (2015) as closely as possible. I flag a share class observation as index fund if contains "INDEX", "ETF", "ISHARES", "IDX", "INDX" in its (uppercase letters) fund name, et_flag is either "F" or "N", the lipper class belongs to S&P500 index ($lipper_class$ is "SPSP" or "SP"), the $index_fund_flag$ is "Y", the FactSet's fund type is either "ETF" or "ETN", the (uppercase letters) brand name is "ISHARES", or it is an index levered fund. I identify a share classes to be of an index levered funds if it has FactSet's leverage factor

larger than 1, or if it contains "INVERSE", "SHORT", "ULTRA", "2X", "3X", "4X", "5X", "6X", "7X", "8X", "9X", "0X", "SHORT TERM", "SHORT TM", "SHORT BOND", "SHORT BND", "LONG SHORT", "LG SHORT" in its (uppercase letters) fund name. I classify Vanguard funds if a fund's name matches one of these (uppercase letters) names: "VFINX", "VEXMX", "NAESX", VEURX", VPACX", "VVIAX", "VBINX", "VEIEX", "VIMSX", "VISGX", or "VISVX". Additionally, following Ben-David et al. (2018), I identify ETFs in my sample merging it with a list of ETFs from CRSP Monthly Stock File ($shrcd$==73). This merge results in 271,198/7,403,148 matches (3.66% of total.)

I classify institutional share classes as observations with $inst_fund$ equal to "Y", or if it contains "INSTITUTIONAL SHARES", "INSTITUTIONAL CLASS", "CLASS I", or "CLASS Y" in its (uppercase letters) fund name. I classify retail share classes as observations with $retail_fund$ equal to "Y", or if it contains "RETAIL SHARES", "RETAIL CLASS", "CLASS A", "CLASS B", "CLASS C", or "INVESTOR CLASS" in its (uppercase letters) fund name.

Then, I aggregate observations across all share classes of the same fund and keep only equity mutual funds and ETFs. I sum the AUM of all share classes, and average all other variables (e.g., expense ratio, returns, turnover, etc.), weighted by lagged AUM. I also keep the first offer date of the oldest share class within fund. Finally, I drop observations before the first time a fund reaches more than $5 million in AUM (in January 2000 dollars) if the fund ever reach that threshold in the sample. I drop observations dated before the fund's first offer date to account for incubation bias (Evans, 2010) and I remove observations with less than 2 years in the full sample (Berk and van Binsbergen, 2015). After this step, I have 1,157,599 fund-month observations.

B.5 Website Technologies

I obtain information on a fund website's technology adoption from BuiltWith. In general, website technologies are defined as tools and services like analytics, payment systems, networking and programming scripts that enhance a website's features. For example, *PayPal Credit* is a technology that enables customers to make buy-now-pay-later payments on a website. Other examples of website technologies are *Google Maps*, *Apple Pay*, and *Shopify*. BuiltWith is a company specialized in website profiling, who sells these data to companies

and consultants. They analyze websites' page code and search for specific patterns that identify the usage of technologies —similar to how a virus scanner searches for pattern in files to identify viruses. The most common patterns they use to identify such technologies are HTML tags, cookies, and Javascript snippets found in a website's source code. BuiltWith continuously crawls websites and analyzes their underlying technologies. The provide a comprehensive database of technologies with installation and (eventual) removal date for millions of websites. They mark a technology as "removed", if they don't find it for two consecutives crawls on a website's code. Appendix Figure B.1 shows a snapshot of the technology data for `arrowfunds.com`, as it appears in my sample. For each unique website in my sample (from CRSP Mutual Funds data) I collect all the technologies name, and installation/removal dates. BuiltWith also provide a $technology_category$ (e.g., Analytics, Feeds, etc.) for each technology, that I map to all technologies installed at least once in my sample. Then, I build a panel with $website$-$technology_name$-month where the first month is the $first_detected$ month, and the last one is the $last_detected$ month. I further filter for analytics' technologies that are aimed to collect and analyze customers' data, and I count the number of such technologies installed in the website. Finally, I merge this dataset by $website$-$date$ to the main CRSP/Morningstar/FactSet data at the share class level (before aggregation). I merge 3,091,431/7,614,808 observations (i.e., 40.60% of the share class-month observations starting March 1993 have at least one data technology in place). Then, I replace missing with zeros, if I have a valid website for the share class-month observation and I have data from BuiltWith for the associated $website$-month, but BuiltWith does not detect data analytics technologies for that month.

Domain	Technology_name	First_Detected	Last_Detected	Live
arrowfunds.com	Ahrefs Bot Disallow	17/09/20	17/07/24	Yes
arrowfunds.com	AJAX Libraries API	14/05/18	24/07/24	Yes
arrowfunds.com	Amazon	21/02/23	13/07/24	Yes
arrowfunds.com	Anthropic Claude Bot Disallow	06/06/24	17/07/24	Yes
arrowfunds.com	ASP.NET	30/05/13	24/07/24	Yes
arrowfunds.com	Baidu Bot Disallow	17/09/20	17/07/24	Yes
arrowfunds.com	Careers	09/03/19	18/07/20	No
arrowfunds.com	Cart Functionality	20/04/23	20/04/23	No
arrowfunds.com	Cohere AI Disallow	18/06/24	17/07/24	Yes
arrowfunds.com	Common Crawl Bot Disallow	06/06/24	17/07/24	Yes
arrowfunds.com	CrUX Dataset	31/12/22	16/07/24	Yes
arrowfunds.com	DoubleClick.Net	26/02/18	24/07/24	Yes
arrowfunds.com	Financial Industry Regulatory Authority	23/12/19	11/08/21	No
arrowfunds.com	Font Awesome	16/04/24	16/04/24	Yes
arrowfunds.com	Global Site Tag	15/02/18	24/07/24	Yes
arrowfunds.com	GoDaddy	02/07/19	13/07/24	Yes
arrowfunds.com	Google AdWords Conversion	15/02/18	05/07/20	No
arrowfunds.com	Google Analytics	21/10/13	24/07/24	Yes

FIGURE B.1: **Example of website technology data**. This figure displays a snapshot of the website technology data from BuiltWith for `arrowfunds.com` (in my sample).

B.6 Final Funds Sample

From the sample of 1,157,599 fund-month observations, I compute fund flows following Lou (2012):

$$Flow_{i,t} = \frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t}) - MGN_{i,t}}{AUM_{i,t-1}}, \tag{B.1}$$

where $AUM_{i,t}$ represents total net assets for fund i in month t, $r_{i,t}$ is the (gross) monthly return, and $MGN_{i,t}$ is the increase in AUM due to the fund's mergers (if any) in month t. Since CRSP does not reports the exact date in which the merger takes place, I follow Lou (2012) and use information about the latest available NAV of the target funds to build a six-months window where the merger plausibly took place. In particular, from CRSP I observe the last date in which the target fund has non-empty NAV, and the identifier of the acquirer. For the acquiring fund, I build a six months window which starts one month before the latest available date of the target fund, until five months after. Within this window, I first compute the flows without accounting for the possible merger (i.e., $\frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1+r_{i,t}}{AUM_{i,t-1}}$) and I flag as $merger_month$ the date with highest flow within the six-months window. Appendix Table B.1 gives an example of this approach for the acquirer fund $crsp_fundno == 662$. In the example, the target fund had latest AUM of $452.5 million (latest date 1999m4). Around the six-months window, the acquirer has one clear $flow$ outlier (computed without accounting for the merger); i.e., a +2,294% net flow in 1995m5. I flag the 1999m5 observation as merger date. Therefore, following equation B.1, the actual fund flow in 1999m5 is -0.6417.

Finally, I keep observations with available variables for my analysis (i.e., $Flow$, AUM, fees). I remove fixed income mutual funds, money market funds, variable products, and others (e.g., 529 Plan, Collective Investment Trust). The monthly sample now contains only ETFs and equity (open-end) mutual funds from March 1993 to December 2023.

$crsp_fundno$	$month$	TNA	$window6M$	$trgt_lastTNA$	$flow$	$flagMerger$
662	1999m1	19.94	0	.	0.0203	0
662	1999m2	18.71	0	.	-0.0127	0
662	1999m3	20.14	1	452.5	0.0517	0
662	1999m4	19.19	1	452.5	-0.0818	0
662	**1999m5**	**459.05**	**1**	**452.5**	**22.9441**	**1**
662	1999m6	450.48	1	452.5	-0.0645	0
662	1999m7	410.50	1	452.5	-0.0478	0
662	1999m8	400.31	1	452.5	-0.0182	0
662	1999m9	368.87	0	.	-0.0510	0
662	1999m10	375.46	0	.	-0.0415	0

TABLE B.1: **Example of Funds Merger**: This table shows an example of funds merger attribution date on CRSP. In this case, the attributed merger month is 1999m5, since it has the largest $flow$ within the six months window around the target latest AUM (1999m4).

B.7 N-SAR Filings

Form N-SAR filings are SEC regulatory files that US registered investment companies (including open-end mutual funds in my sample) used to report until June 1, 2018.[38] N-SAR filings were filed semi-annually and contains a host of detailed information including a fund's service providers, distribution activity, brokerage, and governance. I obtain N-SAR regulatory filings data from Wharton Research Data Service (WRDS) from January 2006, to June 2018. Importantly, each "registrant" (or fund complex) files a separate N-SAR form and reports information for all funds within the fund complex. Each registrant has a unique cik code, but it can include more than one fund. To extract fund-level information in a registrant group, I use three approaches. First, for registrant groups with only one fund the approach is straightforward: I link the information in the Form N-SAR to the (unique) fund reported by the registrant. This step yields 18,843 observations.

Second, for registrants with more than one fund, I first match the registrant group with all funds contained in the complex, by cik-month. Then, I fuzzy match by funds' name ($series_name$) within registrant group using Stata command RECLINK2. I require that the precision of the fuzzy match is above 97.5%. This procedure gives 185,236 perfect matches,

[38] As part of the Investment Company Reporting Modernization initiative, the SEC rescinded Form N-SAR effective June 1, 2018, replacing it with Form N-CEN and Form N-PORT.

and 75,688 fuzzy matches with average precision 99.80%.

Third, for funds without a valid $series_name$, I match to the respective registrant by cik code. Then, for all funds in a fund complex I obtain the turnover and net flows in the N-SAR filings and compute all the pairwise correlations with turnover and net flows from the main CRSP/Morningstar/FactSet dataset (see Appendix B.4) within cik. Finally, I link funds with pairwise correlation of turnover *and* net flows above 99%. After this step, I have 595 additional observations.

B.8 Mutual Fund Prospectus Data

I retrieve mutual funds' prospectuses from the SEC's EDGAR. I focus on Forms 485APOS and 485BPOS, which contain a fund's statutory prospectus. The quarterly sample is from 2006 to 2023. When a mutual fund does not amend its prospectus in each quarter, I consider its last prospectus available. Starting in 2006, the SEC began requiring series ID (fund identifier) and class ID (share class identifier) in each filing. This reporting standard allow to easily link SEC regulatory filings data with major commercial dataset, like CRSP/Morningstar/FactSet. I follow Mullally and Rossi (2025), and use the identifiers' table available at the SEC's website (`https://www.sec.gov/open/datasets-investment_company`) to link mutual funds' prospectuses to the main sample.

For each filing, I download the raw text from EDGAR and parse it to remove HTML tags, XML codes, and other text markups. I then identify candidate Principal Investment Strategy (PIS) sections using regular expressions that search for common section headers such as "Principal Investment Strategy(ies)", "Investment Strategy", or "Primary Strategy", and terminate at typical subsequent sections such as "Principal Risks" or "Fees and Expenses". Once I identify a candidate section, I further screen for its plausibility of being a PIS section in two ways.

First, I exclude spurious matches near the table of contents or summary sections. These PIS candidate are typically short texts (between 5 and 20 words) around the summary section of the fund's prospectus, and clearly represent false positive PIS sections. Second, I require that the section contain characteristic language associated with investment strategies, such as "under normal circumstances...", "the fund seeks..." or "the fund normally invests

primarily. . . ". I clean each section by removing formatting characters, and stripping out numbers unlikely to represent percentages or portfolio weights. Appendix A in (Abis and Lines, 2024, p. 21) shows that the cross sectional distribution of PIS word counts ranges from 20 to around 1,500. Therefore, I remove candidate PIS section with less than 20 words, or excessively long blocks (>2,500 words).

When a filing reports multiple series, I match each PIS to the appropriate series in two steps. I first search for the cleaned series name directly in each PIS, giving priority to longer and more distinctive names to avoid misclassification. For unmatched cases, I then search the surrounding context (up to 10,000 words preceding the section) for the series name and assign the section accordingly. If multiple matches remain, I retain the one with the closest textual proximity with a series ID's name; if no confident match can be made, I flag the filing and match it manually. Each PIS is stored at the series-date level with its respective cik code, series ID, and filing date.

B.9 Mutual Funds Portfolio Holdings

I obtain mutual funds' portfolio holdings data from Thomson Reuters (s12). The database reports fund-level security holdings filed with the SEC. My sample starts in 2004Q2, the first quarter in which funds were required to report holdings quarterly; coverage before then is irregular and incomplete (Harris, Hartzmark and Solomon, 2015). I merge the Thomson Reuters (s12) portfolios with the main CRSP Mutual Funds data using the WRDS MF_LINK tables, which provides a portfolio-level mapping between the two datasets. The final sample consists of quarterly portfolio holdings for US equity mutual funds from 2004Q2 to 2023Q4.

Previous research (e.g., Shive and Yun, 2013) reports a discontinuity in Thomson Reuters (s12) mutual funds holdings coverage after 2008, when compared to CRSP holdings. A common solution in the literature has been to use s12 holdings before June 2008 and CRSP thereafter. However, Thomson Reuters seems to have solved this issue in later vintages (see Appendix Figure B.2). Thus, I prefer not to append different data sources and use the updated s12 holdings throughout the sample.

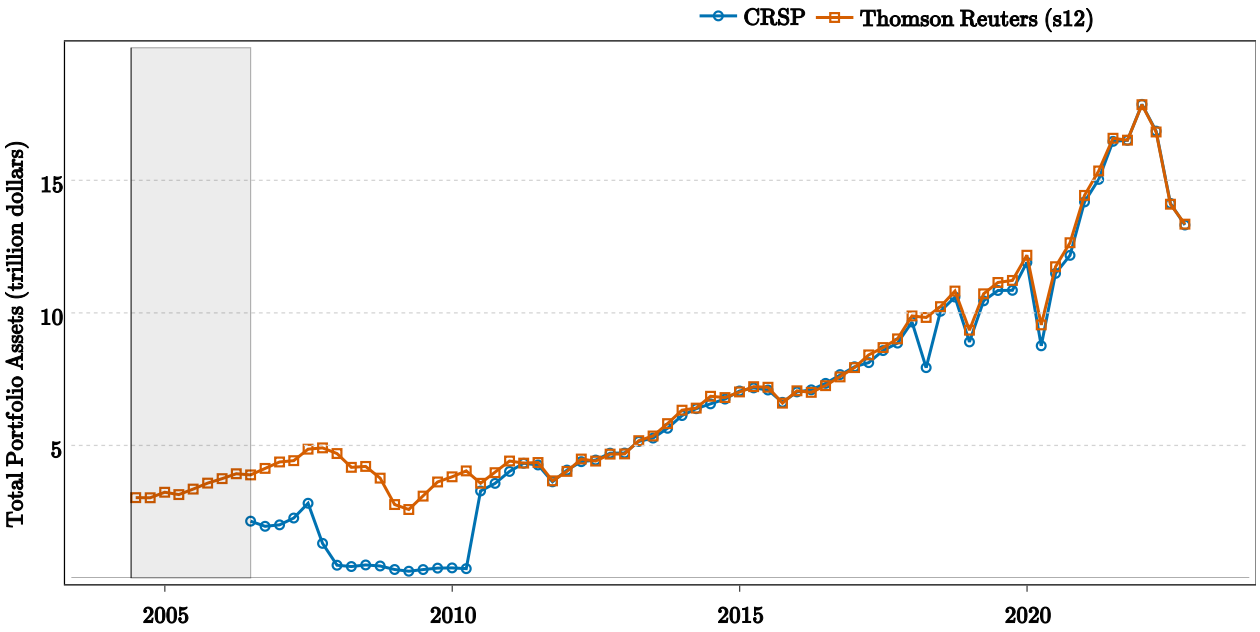


FIGURE B.2: **Total Assets in Thomson Reuters (s12) and CRSP Holdings**. The figure shows the total assets (in trillion of US dollars) in the Thomson Reuters (s12) and CRSP mutual funds holdings databases. CRSP has no holdings data before 2003 (shaded area). The solid blue line reports total equity assets in the Thomson Reuters (s12), while the dotted red line show the CRSP mutual funds holdings data. This figure updates Figure 1 in Shive and Yun (2013) using updated vintages of data.

C List of CIP Codes for Instrumental Variable

CIP Code	Program Title	Broad Category
30.7001	Data Science, General	Data Analytics
30.7099	Data Science, Other	Data Analytics
30.7101	Data Analytics, General	Data Analytics
30.7102	Business Analytics	Data Analytics
30.7103	Data Visualization	Data Analytics
30.7199	Data Analytics, Other	Data Analytics
30.7104	Financial Analytics	Data Analytics
11.0301	Data Processing Technology	Data Analytics
11.0802	Data Modeling/Warehousing and Database Admin.	Data Analytics
27.0501	Statistics, General	Statistics
27.0601	Applied Statistics, General	Statistics
52.1302	Business Statistics	Statistics
27.0502	Mathematical Statistics and Probability	Statistics
27.0503	Mathematics and Statistics	Statistics
27.0599	Statistics, Other	Statistics
11.0501	Computer Systems Analysis/Analyst	Computer Science
11.0401	Information Science/Studies	Computer Science
30.3001	Computational Science	Computer Science
30.0601	Systems Science and Theory	Computer Science
30.0801	Mathematics and Computer Science	Computer Science
11.0101	Computer and Information Sciences, General	Computer Science
11.0701	Computer Science	Computer Science
11.0104	Informatics	Computer Science
11.0103	Information Technology	Computer Science
11.0901	Computer Systems Networking	Computer Science
11.0902	Cloud Computing	Computer Science
11.1001	Network and System Administration	Computer Science
11.1003	CS Security/Information Assurance	Computer Science
11.1002	System, Networking, LAN/WAN Mgmt.	Computer Science
11.1005	IT Project Management	Computer Science
52.1201	Management Information Systems, General	Computer Science

TABLE C.1: **CIP Codes for Data Analytics-Related Fields**: This table reports the Core Instructional Programs (CIP) that I use to construct the local supply of graduates in data analytics. I obtain degree counts from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees awarded by U.S. universities. The annual sample period is from 2000 to 2023.

D Mutual Funds Prospectuses Text Classification

This appendix describes the text classifications and cleaning of mutual funds prospectuses text. I extract the Principal Investment Strategy (PIS) text from SEC Form 485 prospectuses (Form 485APOS and 485BPOS) and the frequency at which specific words and phrase are mentioned. The goal is to capture whether prospectuses emphasize themes that appeal to retail investors (Ben-David et al., 2023).

The unit of observation is the PIS of a mutual fund series in a given fiscal quarter. I obtain all Form 485 filings' plain text from the SEC's EDGAR. I merge fund prospectuses to the CRSP-Morningstar-FactSet dataset using the series ID (fund identifier) and class ID (share class). I keep the last prospectus filing for each series-quarter. This procedure avoids double-counting and accounts for prospectuses' amendments. I lowercase all text, remove HTML tags, formatting, and punctuation. All word and phrase counts are scaled by the total number of words in the PIS.

The resulting measure is expressed per 100 words of text, to facilitate interpretation. I use both tokens precise matches (exact words) and phrase-level matches. Phrase matches allow me to capture bigrams or regular expressions. As I count both token and phrase matches, phrase matches do not double-count overlapping token matches. As robustness checks, I recompute measures under three alternative schemes: (i) excluding stopwords from the denominator, (ii) capping repeated hits within the same sentence at one, and (iii) dropping observations with fewer than 200 words. These alternative definitions do not change the results.

Thematic categories proxy for a fund tends to mention retail-oriented themes in its PIS. I classify words and phrases into themes such as ESG, AI, or clean energy. I select themes based on examples in (Ben-David et al., 2023). Table D.1 provides the full list of tokens and phrases by theme. I compute the composite measure as the sum of all themes mentions.

In addition to theme counts, I compute a text readability index. I measure readability using the Flesch Reading Ease (FRE) index (Flesch, 1948). It is a common measure that summarizes the readability of a text by penalizing long sentences and complex words. Higher values of FRE indicate easier text.

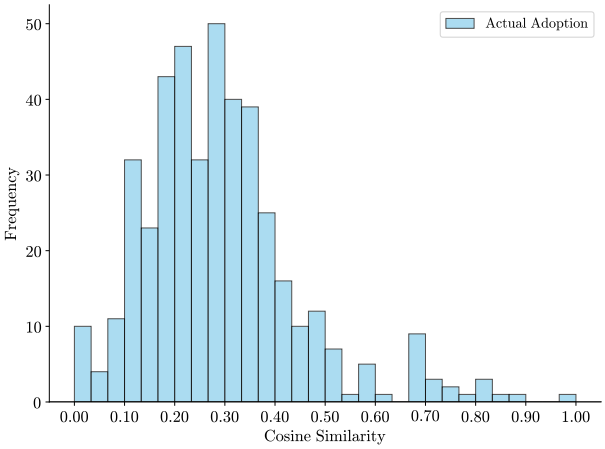
Category	Tokens and Phrases
ESG	esg, sustainable, impact, environmental, governance, responsible
Clean Energy	renewable, solar, wind, battery, ev, "clean energy", "energy transition"
AI	ai, robotics, neural, "artificial intelligence", "machine learning", "generative ai"
Cryptocurrency	crypto, bitcoin, ethereum, blockchain, "digital asset(s)"
Cannabis	cannabis, marijuana, hemp, cbd
Cybersecurity	cybersecurity, infosec, malware, "data breach", "cyber attack"
Religious Values	faith, biblical, christian, islamic, sharia, halal, "faith-based", "values-based"
Political Values	conservative, liberal, republican, democrat, "pro-life", "second amendment"
Space	satellite, rocket, launch, aerospace, "space exploration"
Video Games	gaming, gamers, console, "video game(s)", e-sports

TABLE D.1: **Themes Classification and Words/Phrases**: This table reports the list of words and sentences associated with each retail-oriented theme. I count the number of word phrase matches in each fund PIS, scale by the total number of words, and express the final measure per 100 words of text.

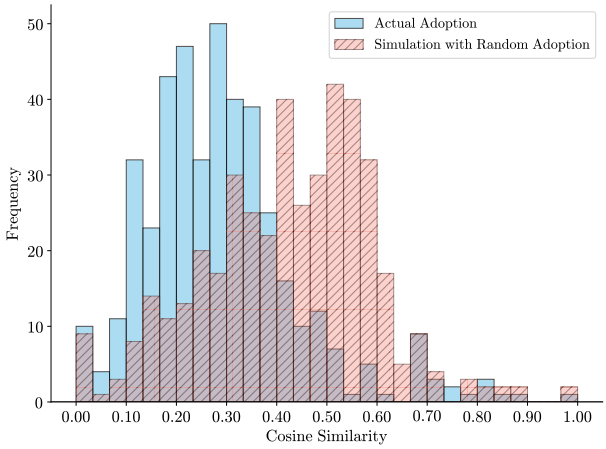
E Additional Results

This section contains additional results and robustness not contained in the main text.

E.1 Appendix Figures



(A:) Overlap in Technology Adoption.



(B:) Comparison with Random Simulation.

FIGURE E.1: **Overlap in Technology Adoption within Hosting Platform**. The figure shows the cosine similarity in website technology adoption within website hosting platforms. Cosine similarity is higher when there is high overlap among website technologies within a given hosting platform. Panel A shows the histogram of similarities for the sample of US mutual funds in this paper (mean 0.29, median 0.27). Panel B also reports, in red, the histogram of similarities for counterfactual simulations where technology adoption within hosting platforms is random (mean 0.43, median 0.44).



FIGURE E.2: **Word Cloud of Data Technologies' Descriptions**. The figure shows the word cloud of data technologies' features in my sample. For each data technology, I collect a short description of what the technology allows to do (e.g., ipstack: *"provides IP to geolocation APIs and global IP database services*") and plot the word cloud where larger fonts indicate a higher word frequency.

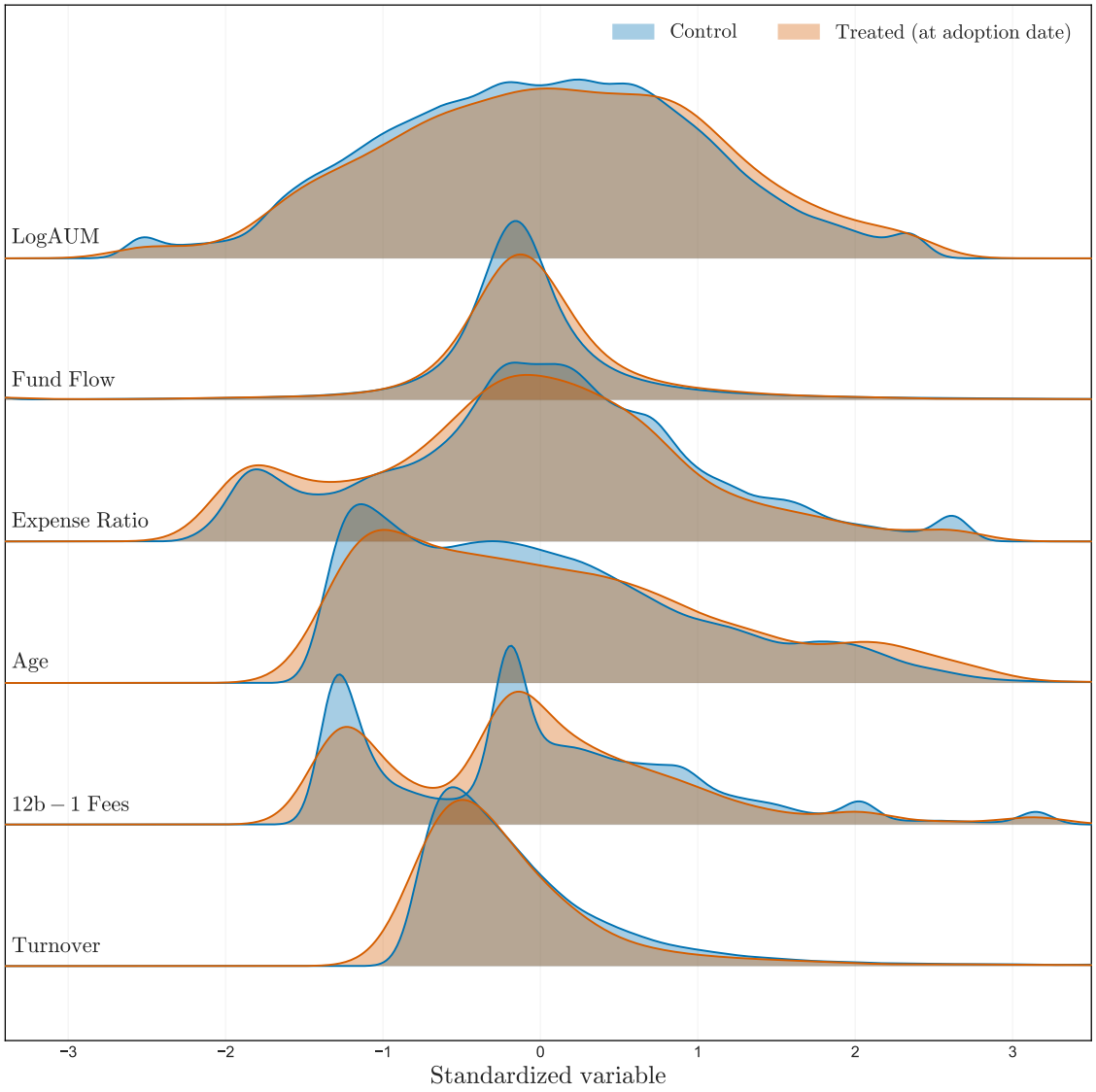


FIGURE E.3: **Balance Covariates**. This figure shows the empirical pdf of covariates for funds in the treatment and control group. All variables are normalized to have a mean of zero and a standard deviation of one in the full sample. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t. For each group, I report the covariates in the 3 months pre-adoption. The figure shows covariates included in regressions in the main text.

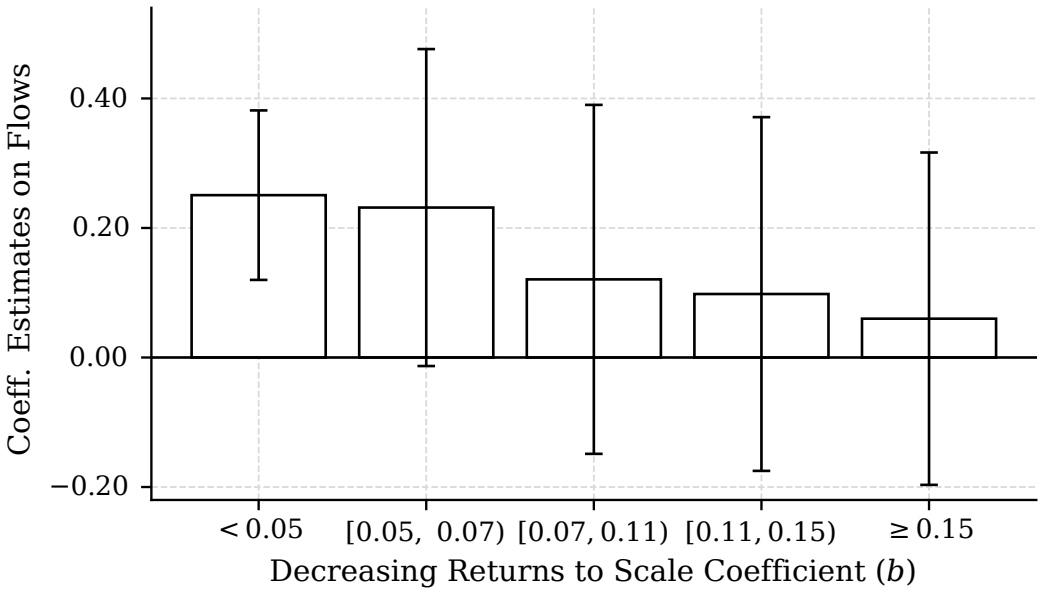


FIGURE E.4: **Heterogeneity of Effect on Flows by Decreasing Returns to Scale**. This figure shows results for difference-in-differences coefficients across different values of decreasing returns to scale (b) as of data technology adoption. I estimate decreasing returns to scale as-of adoption following the recursive demeaning approach proposed in Pástor, Stambaugh and Taylor (2015). The coefficient b captures the extent to which a fund's size reduces its ability to generate alpha. Each bar represents a level of subsample by decreasing returns to scale coefficient (e.g., the first bar represents all funds with almost negligible decreasing returns to scale). Each vertical line represents the 95% confidence interval. The specification is the same as the main specification in equation (2). All regressions include fund and category×time fixed effects, and controls: fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t.

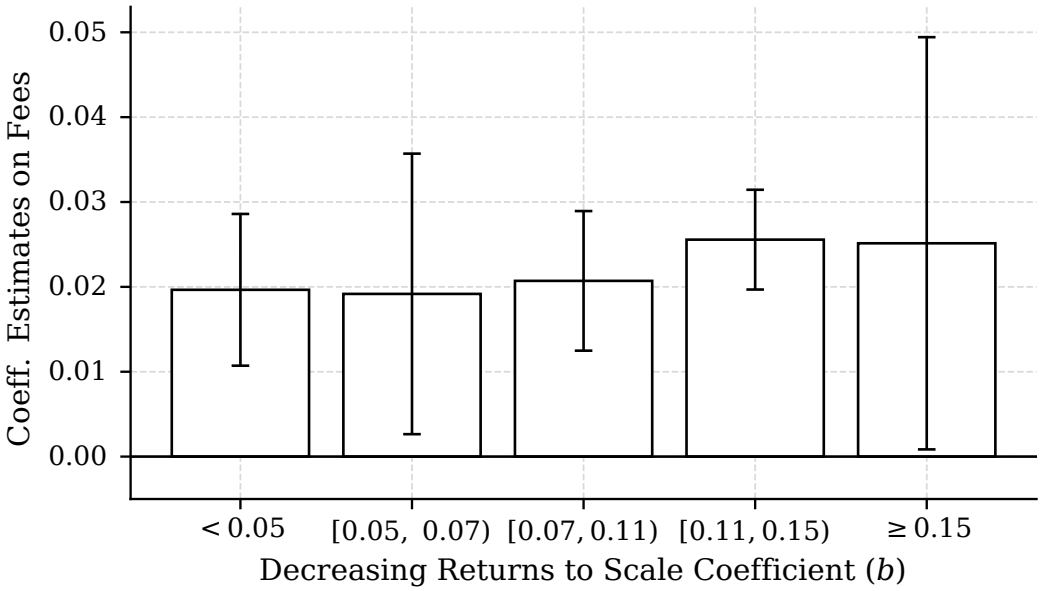


FIGURE E.5: **Heterogeneity of Effect on Expense Ratio by Decreasing Returns to Scale**. This plot report a placebo test on the heterogeneity of the effect of data technology on expense ratio by subsample based on decreasing returns to scale The figure shows results for difference-in-differences coefficients across different values of decreasing returns to scale (b) as of data technology adoption. I estimate decreasing returns to scale as-of adoption following the recursive demeaning approach proposed in Pástor, Stambaugh and Taylor (2015). The coefficient b captures the extent to which a fund's size reduces its ability to generate alpha. Each bar represents a level of subsample by decreasing returns to scale coefficient (e.g., the first bar represents all funds with almost negligible decreasing returns to scale). Each vertical line represents the 95% confidence interval. The specification is the same as for Table 5. All regressions include fund and category×time fixed effects, and controls: fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t.

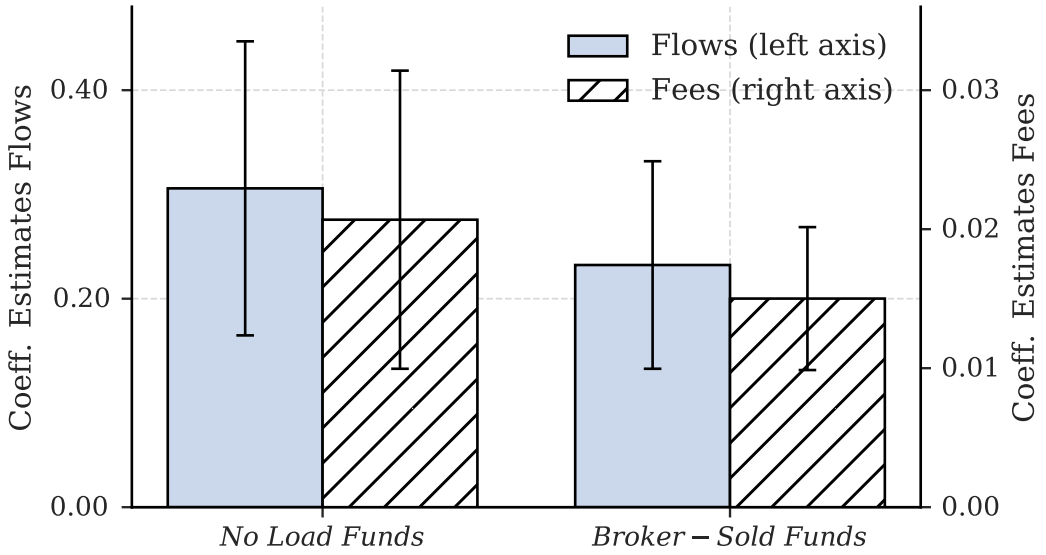


FIGURE E.6: **Direct Sold vs Broker-Sold Funds**. This figure shows results for difference-in-differences coefficients for no-load funds and broker-sold funds. No-load funds are funds that are directly sold to investors Del Guercio and Reuter (2014). I identify directly sold (no-load) funds and broker-sold funds following (Sun, 2021). Each bar represents a coefficient estimates on fund flows (plain bars) or expense ratio (hatched bars). The left (right) axis refers to flows (fees). Vertical line represents the 95% confidence interval. The specification is the same as the main specification in equation (2). All regressions include fund and category×time fixed effects, and controls: fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t.

E.2 Appendix Tables

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$		0.116***		0.101**
		(0.044)		(0.043)
logAUM$_{i,t}$	-0.420***	-0.421***	-0.420***	-0.421***
	(0.025)	(0.025)	(0.025)	(0.025)
logAge$_{i,t}$	-1.889***	-1.889***	-1.871***	-1.871***
	(0.068)	(0.068)	(0.066)	(0.066)
CAPM Alpha$_{i,t}$	6.749***	6.749***	9.185***	9.182***
	(0.348)	(0.348)	(0.457)	(0.457)
Turnover$_{i,t}$	0.025	0.026	0.027	0.027
	(0.040)	(0.040)	(0.036)	(0.036)
12b-1 Fees$_{i,t}$	-0.156	-0.150	-0.115	-0.109
	(0.101)	(0.101)	(0.099)	(0.099)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266
Obs.	947,079	947,079	946,733	946,733
Adj. R^2	0.094	0.094	0.126	0.126

TABLE E.1: **Fund Flows and Data Technologies, detailed covariates results**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics. See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Pre-adoption growth rates	Control	Treatment	Difference	p-value
Monthly Fund Flows	-0.192	-0.201	0.008	0.785
Quarterly Fund Flows	-0.250	-0.287	0.038	0.274

TABLE E.2: **Parallel trends**: This table reports the growth rate of fund flows in the 12 months pre-adoption, for adopting and not-adopting fund. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t. The table shows monthly and quarterly growth rates of fund flows for the sample of treated and control group in the 12 months pre-adoption. I winsorize growth rates at the 1% and 99% level. The last column reports the p-value of the difference between treated and control groups.

Panel A: Treated (at adoption)							
	mean	sd	p5	p25	p50	p75	p95
AUM ($M)	1,137.59	2,786.38	7.25	45.16	196.86	848.55	5,757.72
Expense Ratio (%)	1.06	0.53	0.13	0.73	1.07	1.39	1.97
12b-1 Fees (%)	0.28	0.21	0.00	0.10	0.25	0.39	0.71
Flows (%)	0.30	6.06	-5.54	-1.55	-0.41	1.14	8.24
Turnover Ratio	0.75	1.08	0.05	0.23	0.46	0.85	2.31
Age (Years)	12.48	8.76	1.42	5.08	11.08	18.00	29.75
Panel B: Control							
	mean	sd	p5	p25	p50	p75	p95
AUM ($M)	930.97	2,440.76	6.69	39.78	169.32	665.31	4,227.31
Expense Ratio (%)	1.12	0.53	0.18	0.80	1.13	1.44	2.01
12b-1 Fees (%)	0.29	0.22	0.00	0.10	0.26	0.43	0.74
Flows (%)	0.20	6.12	-5.77	-1.77	-0.57	1.00	8.76
Turnover Ratio	0.83	1.03	0.07	0.27	0.54	0.99	2.40
Age (Years)	11.58	8.10	1.25	4.83	10.25	16.50	27.50

TABLE E.3: **Covariates balance.**: This table reports summary statistics of covariates for funds in the treatment and control group. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t. For each group, I report the covariates in the 3 months pre-adoption. The table shows covariates included in regressions in the main text. AUM is inflation adjusted in January 2000 $ million. Expense Ratio, 12b-1 Fees, and Flows are in %; e.g., the average fund flow for the control group is 0.20% monthly.

	Fund Flows$_{i,t+1}$ (%)							
	(1)	(2)	(3)	(4)	(5)	(6)	(6)	(8)
Data$_{i,t}$	0.132***	0.118***	0.115***	0.101**	0.178***	0.168***	0.157***	0.145***
	(0.040)	(0.040)	(0.043)	(0.043)	(0.030)	(0.045)	(0.056)	(0.043)
Estimator	OLS	OLS	OLS	OLS	Stag-DiD	Stag-DiD	Stag-DiD	Stag-DiD
Baseline controls:	✓	✓	✓	✓	✓	✓	✓	✓
Additional controls:								
Expense Ratio $_{i,t}$	✓	✓	×	×	✓	✓	×	×
Fund Flows$_{i,t}$	✓	✓	×	×	✓	✓	×	×
Morningstar Rating$_{i,t}$	×	×	✓	✓	×	×	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266	6.266	6.266	6.266	6.266
Obs.	947,079	946,733	947,079	946,733	890,800	873,140	890,801	873,141
Adj. R^2	0.108	0.138	0.100	0.132	0.108	0.138	0.100	0.132

Table E.4: **Fund Flows and Data Technologies, additional controls**: This table shows results of panel regression robust to several additional controls in addition to the baseline regression. The dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. In columns (1), (2), (5), and (6) I also include past fund flow and expense ratio as controls. Columns (3), (4), (7), and (8) add the Morningstar Rating in the set of controls. Columns (1) to (4) report results for baseline OLS, while columns (5) to (8) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	logAUM$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.114***	0.102***	0.117***	0.123***
	(0.022)	(0.022)	(0.028)	(0.030)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	5.216	5.216	5.216	5.216
Outcome SE	2.007	2.007	2.007	2.007
Obs.	947,323	946,732	890,759	873,078
Adj. R^2	0.842	0.854	0.842	0.854

TABLE E.5: **Fund Size and Data Technologies**: This table shows results of panel regression in which the dependent variable is the fund size –that is, a fund's (log) AUM. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. I exclude the lagged fund's size as a control, because it might introduce post-treatment bias (Roberts and Whited, 2013). Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)							
	(1)	(2)	(3)	(4)	(5)	(6)	(6)	(8)
Data$_{i,t}$	0.001*	0.001*	0.019***	0.018***	0.001***	0.001***	0.037***	0.034***
	(0.000)	(0.000)	(0.004)	(0.004)	(0.000)	(0.000)	(0.005)	(0.005)
Estimator	OLS	OLS	OLS	OLS	Stag-DiD	Stag-DiD	Stag-DiD	Stag-DiD
Baseline controls:	✓	✓	✓	✓	✓	✓	✓	✓
Additional controls:								
Expense Ratio $_{i,t}$	✓	✓	×	×	✓	✓	×	×
Fund Flows$_{i,t}$	✓	✓	×	×	✓	✓	×	×
Morningstar Rating$_{i,t}$	×	×	✓	✓	×	×	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	1.133	1.133	1.133	1.133	1.133	1.133	1.133	1.133
Outcome SE	0.540	0.540	0.540	0.540	0.540	0.540	0.540	0.540
Obs.	947,079	946,733	947,079	946,733	890,801	873,141	890,802	873,141
Adj. R^2	0.997	0.997	0.915	0.919	0.997	0.997	0.915	0.919

Table E.6: **Expense Ratio and Data Technologies, additional controls**: This table shows results of panel regression robust to several additional controls in addition to the baseline regression. The dependent variable is the one-month-ahead fund expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. In columns (1), (2), (5), and (6) I also include past fund flow and expense ratio as controls. Columns (3), (4), (7), and (8) add the Morningstar Rating in the set of controls. Columns (1) to (4) report results for baseline OLS, while columns (5) to (8) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Data$_{i,t}$	0.137***	0.124**	0.144**	0.150***	0.139***	0.147***
	(0.045)	(0.057)	(0.058)	(0.043)	(0.054)	(0.047)
Baseline controls:	✓	✓	✓	✓	✓	✓
Performance control:						
Vanguard Alpha	✓	✓	×	×	×	×
FF 3-Factor Alpha	×	×	✓	✓	×	×
FF 5-Factor Alpha	×	×	×	×	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266	6.266	6.266
Obs.	815,786	796,957	892,004	874,345	892,003	874,344
Adj. R^2	0.094	0.126	0.089	0.118	0.080	0.110

Table E.7: **Fund Flows and Data Technologies using different measures of performance**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The baseline control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees. Columns (1) and (2) additionally control for alpha with respect to a set of passive Vanguard funds, following Berk and van Binsbergen (2015). Columns (3) and (4) use Fama and French (1993) 3-Factors, while columns (5) and (6) use alpha with respect to the Fama and French (2015) 3-Factors model. All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
DATA$_{i,t}$	0.037***	0.034***	0.037***	0.034***	0.037***	0.034***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Baseline controls:	✓	✓	✓	✓	✓	✓
Performance control:						
Vanguard Alpha	✓	✓	×	×	×	×
FF 3-Factor Alpha	×	×	✓	✓	×	×
FF 5-Factor Alpha	×	×	×	×	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	1.133	1.133	1.133	1.133	1.133	1.133
Outcome SE	0.540	0.540	0.540	0.540	0.540	0.540
Obs.	890,802	873,141	892,004	874,345	892,004	874,345
Adj. R^2	0.915	0.919	0.922	0.926	0.922	0.926

TABLE E.8: **Expense Ratio and Data Technologies using different measures of performance**: This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The baseline control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees. Columns (1) and (2) additionally control for alpha with respect to a set of passive Vanguard funds, following Berk and van Binsbergen (2015). Columns (3) and (4) use Fama and French (1993) 3-Factors, while columns (5) and (6) use alpha with respect to the Fama and French (2015) 3-Factors model. All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)		Expense Ratio$_{i,t+1}$ (%)	
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.275***	0.188**	0.061***	0.052***
	(0.053)	(0.088)	(0.005)	(0.006)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.188	0.188	1.133	1.133
Outcome SE	6.266	6.266	0.540	0.540
Obs.	890,802	873,141	890,802	873,141
Adj. R^2	0.094	0.124	0.922	0.926

Table E.9: **Robustness of Main Results excluding Google Analytics**: This table addresses concerns that the results are entirely driven by Google Analytics. The table replicates the main findings in Tables 3 and 5 excluding Google Analytics from the set of data technologies. Columns (1) and (2) show results for flows, while columns (3) and (4) report the robustness for results on expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.116*	0.101*	0.148***	0.171***
	(0.059)	(0.059)	(0.053)	(0.057)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.187	0.187	0.187	0.187
Outcome SE	6.264	6.264	6.264	6.264
Obs.	947,132	946,599	890,678	873,040
Adj. R^2	0.094	0.126	0.094	0.126

Table E.10: **Robustness Results on Flows, clustering by Fund Family**: This table shows robustness results for the main specification, clustering standard errors by fund family. The dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.017***	0.016***	0.037***	0.034***
	(0.006)	(0.006)	(0.005)	(0.005)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.133	1.133	1.133	1.133
Outcome SE	0.540	0.540	0.540	0.540
Obs.	947,079	946,510	890,678	873,040
Adj. R^2	0.922	0.926	0.922	0.926

Table E.11: **Robustness Results on Expense Ratio, clustering by Fund Family**: This table shows robustness results for the main specification, clustering standard errors by fund family. The dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.223	0.375*	0.329**	0.493*
	(0.217)	(0.202)	(0.139)	(0.282)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.289	0.289	0.289	0.289
Outcome SE	6.074	6.074	6.074	6.074
Obs.	13,276	11,712	12,047	9,011
Adj. R^2	0.120	0.149	0.120	0.149

TABLE E.12: **Robustness for Fund Families with only one website for each fund**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow, for a subset of the main sample. I keep only fund families with a unique website associated to a given fund, and run the main specification. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Family Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.430	0.420	0.816***	0.832***
	(0.448)	(0.448)	(0.264)	(0.232)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.473	0.473	0.473	0.473
Outcome SE	55.606	55.606	55.606	55.606
Obs.	171,010	171,010	154,851	154,851
Adj. R^2	0.007	0.008	0.008	0.008

TABLE E.13: **Robustness aggregating Flows by Fund Family**: This table shows results of panel regression in which the dependent variable is the one-month-ahead within fund family total flow. The regressors are a dummy equal to one if a fund family i has at least a data technology in place at month t (DATA$_{i,t}$), and controls for fund family-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund family's total size (log total AUM), (log) age, and the number of funds in the family in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
Data$_{i,t}$	0.008***	0.006	0.029***	0.026***
	(0.001)	(0.006)	(0.006)	(0.007)
Estimator	OLS	OLS	Staggered DiD	Staggered DiD
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.315	1.315	1.315	1.315
Outcome SE	0.522	0.522	0.522	0.522
Obs.	422,280	421,480	368,724	357,947
Adj. R^2	0.918	0.922	0.918	0.922

Table E.14: **Robustness Results on Expense Ratio, Retail Share Classes Only**: This table shows robustness results for the main specification on expense ratio, for retail share classes only. For each fund, I aggregate only retail share classes and remove all institutional share classes. Then, I run the main specification as in Table 5. The dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$\mathbb{P}$ {Adoption} (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Category Adoption $\%_{i,t}$	0.144	0.153	-0.261	0.131	0.184	0.272
	(0.900)	(0.290)	(0.739)	(0.220)	(0.501)	(0.170)
State Adoption $\%_{i,t}$	2.966***	1.112***				
	(0.817)	(0.275)				
City Adoption $\%_{i,t}$			3.937***	1.404***		
			(0.700)	(0.224)		
Zip Code Adoption $\%_{i,t}$					4.767***	1.894***
					(0.462)	(0.183)
$log\text{AUM}_{i,t}$	0.047*	0.015	0.031	0.008	0.074***	0.022**
	(0.025)	(0.009)	(0.025)	(0.010)	(0.025)	(0.011)
$log\text{Age}_{i,t}$	-0.070	-0.019	-0.063	-0.014	-0.080	-0.020
	(0.052)	(0.019)	(0.053)	(0.020)	(0.049)	(0.021)
Estimator	Logit	Probit	Logit	Probit	Logit	Probit
Obs.	742,516	742,516	742,516	742,516	742,516	742,516
Pseudo R^2	0.087	0.090	0.132	0.131	0.227	0.219

TABLE E.15: **Technology Diffusion in the Asset Management Industry**: This table shows results of logit/probit regression of probability to adopt data technology, on the (lagged) adoption rate at different levels of aggregation. The adoption rate is defined as the percentage of funds with data technology in place, within a given category, state, city, or zip code in month t. Columns (1) and (2) use adoption rate at the state level, columns (3) and (4) at city, and columns (5) and (6) at the zip code level. The regressors are adoption rate within fund category, adoption rate at the geographical level (state, city, or zip code), and controls for fund-month characteristics (omitted for brevity). The control variables include (lagged) fund's size (logAUM) and (log) age. All standard errors are clustered by month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Excluded:	Boston	Chicago	New York	Los Angeles	Philadelphia	San Francisco
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{DATA}_{i,c,y}$	1.559**	1.576*	1.210**	1.813**	1.803***	2.050**
	(0.745)	(0.809)	(0.604)	(0.756)	(0.681)	(1.012)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Category×Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	-1.723	-1.723	-1.723	-1.723	-1.723	-1.723
Outcome SE	17.397	17.397	17.397	17.397	17.397	17.397
Obs.	33,537	37,600	30,961	39,067	38,728	37,550
Adj. R^2	0.322	0.326	0.324	0.326	0.325	0.321
F-Stat	29.268	29.268	29.268	29.268	29.268	29.268

TABLE E.16: **Robustness IV Estimates on Flows: Excluding Financial Districts**: This table shows robustness results for the instrumental variable estimates where the instrument is the local supply of talent in data analytics-related fields. The dependent variable is the one-month-ahead fund flow. In column (1) to (6), I remove funds located in each financial district as defined in Christoffersen and Sarkissian (2009), namely, Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y, with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix C for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y ($\text{DATA}_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in year y. All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Excluded:	Boston (1)	Chicago (2)	New York (3)	Los Angeles (4)	Philadelphia (5)	San Francisco (6)
$\text{DATA}_{i,c,y}$	0.037***	0.034***	0.025***	0.030***	0.032***	0.028***
	(0.008)	(0.008)	(0.010)	(0.009)	(0.012)	(0.005)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Category×Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	1.084	1.084	1.084	1.084	1.084	1.084
Outcome SE	0.527	0.527	0.527	0.527	0.527	0.527
Obs.	33,536	37,600	31,597	39,865	38,727	37,550
Adj. R^2	0.658	0.659	0.651	0.659	0.659	0.657
F-Stat	29.268	29.268	29.268	29.268	29.268	29.268

TABLE E.17: **Robustness IV Estimates on Fees: Excluding Financial Districts**: This table shows robustness results for the instrumental variable estimates where the instrument is the local supply of talent in data analytics-related fields. The dependent variable is the one-month-ahead expense ratio. In column (1) to (6), I remove funds located in each financial district as defined in Christoffersen and Sarkissian (2009), namely, Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y, with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix C for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y ($\text{DATA}_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in year y. All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Expense Ratio$_{i,t+1}$ (%)					
			z_i : Tenure of Adoption		N. of Data Tech.	
	(1)	(2)	(3)	(4)	(5)	(6)
Data$_{i,t}$	0.018**	0.019**	0.010	0.009	0.012	0.010
	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)
Data$_{i,t} \times Post_t$	0.021***	0.022***				
	(0.008)	(0.007)				
Data$_{i,t} \times Post_t \times z_i$			0.000	0.000	0.000	0.000
			(0.002)	(0.002)	(0.002)	(0.002)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	1.174	1.174	1.174	1.174	1.174	1.174
Outcome SE	0.542	0.542	0.542	0.542	0.542	0.542
Obs.	807,811	807,180	584,569	583,935	692,109	691,474
Adj. R^2	0.921	0.925	0.917	0.921	0.928	0.931

Table E.18: **Fund Fees and TensorFlow**: This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. Columns (1) and (2) follow specification in equation (3), while columns (3) to (6) follow (4). In columns (3) and (4) the continuous treatment z_i is the (log) number of months between the first data technology adoption and TensorFlow's release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow's release, as continuous treatment z_i. Data$_{i,t}$ is a dummy equal to one if fund i has a data technology in place at month t. See Section 3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (6)) in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow's release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)					
z_i :			Tenure of Adoption		N. of Data Tech.	
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Data}_{i,t}$	0.567***	0.595***	0.517***	0.517***	0.491***	0.491***
	(0.139)	(0.140)	(0.146)	(0.146)	(0.136)	(0.136)
$\text{Data}_{i,t} \times Post_t$	0.253**	0.313***				
	(0.109)	(0.109)				
$\text{Data}_{i,t} \times Post_t \times z_i$			0.080**	0.080**	0.023**	0.023**
			(0.033)	(0.033)	(0.011)	(0.011)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	0.182	0.182	0.182	0.182	0.182	0.182
Outcome SE	6.458	6.458	6.458	6.458	6.458	6.458
Obs.	598,363	597,524	438,485	438,485	519,866	519,866
Adj. R^2	0.104	0.140	0.093	0.093	0.101	0.101

TABLE E.19: **Robustness on TensorFlow, Excluding Growth Funds**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. I exclude growth funds from this sample to mitigate concerns that results are driven by an increase in expected cash flows for tech stocks. Columns (1) and (2) follow specification in equation (3), while columns (3) to (6) follow (4). In columns (3) and (4) the continuous treatment z_i is the (log) number of months between the first data technology adoption and TensorFlow's release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow's release, as continuous treatment z_i. $\text{Data}_{i,t}$ is a dummy equal to one if fund i has a data technology in place at month t. See Section 3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund's size (logAUM), (log) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (6)) in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow's release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)	
	(1)	(2)
Data$_{i,t}$	0.437***	0.379***
	(0.141)	(0.132)
Data$_{i,t} \times \gamma_{c,t}$	-0.571**	-0.493**
	(0.224)	(0.212)
$\gamma_{c,t}$	0.620	0.021
	(0.422)	(1.607)
logAUM$_{i,t}$	-0.422***	-0.421***
	(0.025)	(0.025)
logAge$_{i,t}$	-1.888***	-1.870***
	(0.068)	(0.066)
CAPM Alpha$_{i,t}$	6.751***	9.183***
	(0.348)	(0.457)
Turnover$_{i,t}$	0.025	0.027
	(0.040)	(0.036)
12b-1 Fees$_{i,t}$	-0.160	-0.117
	(0.101)	(0.099)
Fund FE	✓	✓
Time FE	✓	×
Category×Time FE	×	✓
Outcome mean	0.187	0.187
Outcome SE	6.264	6.264
Obs.	947,307	946,733
Adj. R^2	0.094	0.126

Table E.20: **Fund Flows and Competition**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics. See Section 3.2 for details on data technologies. I include an interaction term with the competition coefficient $\gamma_{c,t}$ for fund category c in month t. The competition coefficient is built following equation (6), and it captures the fraction of funds with data technologies in place within fund category-month. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
N. $Tech_{i,t}$	0.034***	0.049***		
	(0.007)	(0.018)		
N. $Tech^2_{i,t}$		-0.001		
		(0.001)		
log(1+ N. $Tech_{i,t}$)			0.133***	0.131***
			(0.032)	(0.031)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	×	×	✓	×
Category×Time FE	✓	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266
Obs.	798,041	798,041	798,559	798,041
Adj. R^2	0.126	0.126	0.094	0.126

TABLE E.21: **Fund Flows and Number of Technologies**: This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are the number of data technologies in place for fund i at month t (N. $Tech_{i,t}$) in column (1); column (2) adds its square (N. $Tech^2_{i,t}$), and the log of (1+N. $Tech_{i,t}$) in columns (3)-(4). See Section 3.2 for details on data technologies. All columns include controls for fund-month characteristics (omitted for brevity). The control variables are a fund's size (logAUM), (log) age, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Vanguard alpha (%)		CAPM alpha (%)		FF-3 alpha (%)		FF-5 alpha (%)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\text{DATA}_{i,t}$	0.021		0.027*		0.014		0.002	
	(0.016)		(0.015)		(0.015)		(0.022)	
$\overline{\text{DATA}}_{i,t}$		-0.002		0.013		0.025		0.022
		(0.019)		(0.017)		(0.019)		(0.027)
$log\text{AUM}_{i,t}$	-0.159***		-0.134***		-0.103***		-0.036	
	(0.020)		(0.017)		(0.019)		(0.031)	
$\overline{log\text{AUM}}_{i,t}$		-0.087*		-0.048		0.020		0.097
		(0.047)		(0.039)		(0.038)		(0.059)
Estimator	OLS	RD	OLS	RD	OLS	RD	OLS	RD
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓
Outcome mean	-0.651	-0.651	-0.146	-0.146	-0.539	-0.539	-0.449	-0.449
Outcome SE	3.191	3.191	2.993	2.993	4.994	4.994	7.335	7.335
Obs.	971,918	971,918	971,918	971,918	971,918	971,918	971,918	971,918
Adj. R^2	0.174	0.002	0.402	0.001	0.702	-0.001	0.750	-0.001
First stage F-stat	–	5,273.077	–	5,410.182	–	5,417.647	–	5,417.647

TABLE E.22: **Performance and Data Technologies**: This table shows results of regression in which the dependent variable is the one-month-ahead fund performance. Columns (1), (3), (5), and (7) report results for OLS regressions, while columns (2), (4), (6) and (8) use the recursive demeaning approach (RD) in Pástor, Stambaugh and Taylor (2015) which accounts for the positive contemporaneous correlation between fund size and unexpected returns. The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($\text{DATA}_{i,t}$), 12b-1 fees, and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, and turnover. I use Vanguard alpha (columns (1) and (2)), CAPM alpha (columns (3) and (4)), FF-3 factors alpha (columns (5) and (6)), and FF-5 factors alpha (columns (7) and (8)) as proxy of funds' performance. The Vanguard alpha is the risk adjusted performance with respect to the set of available Vanguard index funds (Berk and van Binsbergen, 2015). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Payment Redistribution: (in %)	BD	Captive	Own	Adv.	Prospect
	(1)	(2)	(3)	(4)	(5)
$\text{DATA}_{i,t}$	-3.039***	0.088***	-0.368	-0.061***	0.015*
	(0.806)	(0.030)	(0.779)	(0.023)	(0.008)
Controls	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓
Category×Time FE	✓	✓	✓	✓	✓
Outcome mean	27.031	0.170	40.160	0.488	0.053
Outcome SE	42.480	1.337	48.222	3.180	0.355
Obs.	209,530	209,268	209,253	209,535	209,367
Adj. R^2	0.869	0.760	0.881	0.736	0.641

TABLE E.23: **Redistribution of 12b-1 Fees Payments**: This table shows results of panel regression on funds' distribution to 12b-1 payment categories (in %), robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). The dependent variable is the fund's one-month-ahead percentage allocation of 12b-1 fees in each category. Column (1) report results for the percentage payment towards broker and dealers, column (2) to captive retail sales force, column (3) to the underwriter itself (retained), column (4) to external advertising, and column (5) to printing and mailing of prospectuses to prospect clients. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($\text{DATA}_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 3.2 for details on data technologies. The control variables include a fund's size (logAUM), (log) age, turnover, 12b-1 fees, and CAPM alpha in month t. The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)							
	(1)	(2)	(3)	(4)	(5)	(6)	(6)	(8)
CAPM Alpha$_{i,t}$	0.084***	0.108***						
	(0.004)	(0.004)						
CAPM Alpha$_{i,t}$×Data$_{i,t}$	-0.016***	-0.014***						
	(0.004)	(0.004)						
Vanguard Alpha$_{i,t}$			0.735***	0.988***				
			(0.041)	(0.052)				
Vanguard Alpha$_{i,t}$×Data$_{i,t}$			-0.225***	-0.250***				
			(0.455)	(0.466)				
FF3 Alpha$_{i,t}$					0.058***	0.078***		
					(0.003)	(0.004)		
FF3 Alpha$_{i,t}$×Data$_{i,t}$					-0.006***	-0.004**		
					(0.002)	(0.002)		
FF5 Alpha$_{i,t}$							0.020***	0.025***
							(0.002)	(0.003)
FF5 Alpha$_{i,t}$×Data$_{i,t}$							0.000	0.001
							(0.001)	(0.001)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓	×	✓
Outcome mean	0.188	0.188	0.188	0.188	0.188	0.188	0.188	0.188
Outcome SE	6.266	6.266	6.266	6.266	6.266	6.266	6.266	6.266
Obs.	947,079	946,733	947,078	946,732	890,800	873,140	890,801	873,141
Adj. R^2	0.094	0.124	0.094	0.126	0.087	0.118	0.079	0.110

Table E.24: **Flow-Performance Sensitivity and Data Technologies**: This table shows results of panel regression of flow-performance sensitivity and interaction terms with the adoption of a data technology. The dependent variable is the one-month-ahead fund flow. The regressors are the fund's performance in month t, and interaction term with a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), and controls for fund-month characteristics. The control variables include a fund's size (logAUM), (log) age, turnover, and 12b-1 fees in month t. The first row reports results for the flow-performance sensitivity (Chevalier and Ellison, 1997). The second row shows results for the interaction term of flow-performance sensitivity with the Data dummy. Columns (1)-(2) use CAPM alpha as measure of a fund's performance, columns (3)-(4) use the alpha with respect to a set of 11 existing Vanguard funds (Berk and van Binsbergen, 2015), columns (5)-(6) use Fama-French 3-factors alpha, and columns (7)-(8) use Fama-French 5-factors alpha. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.