Asset (and Data) Managers

Marco Zanotti*
Swiss Finance Institute, USI Lugano

This version: March 14, 2025

Abstract

This paper studies the *direct* impact of new technologies on the asset management industry. I show that technological innovations substantially improve fund managers' ability to target customer demand and attract capital inflows, with implications for the industry's structure. Exploiting information from their websites' codes, I track when fund managers start collecting and analyzing customers' data using tools like Google Analytics. Funds adopting such technologies attract 1.5% higher annual flows, with the effect being concentrated in retail share classes. Additionally, they expand product offerings and charge higher fees. The effects decrease with competition as more funds within the same category adopt similar technologies. Overall, these results show that technological innovation in asset management extends beyond portfolio allocation decisions to impact how funds attract and retain capital. This evidence highlights the economic importance of managers learning from investors' data.

Keywords: Asset Management, Information, Data Economy, FinTech, Big Data

JEL Classification: G14, G23, L10

*Swiss Finance Institute, USI Lugano. Email: marco.zanotti@usi.ch. I am very grateful to Laurent Frésard, Alberto Plazzi, and Andrea Tamoni for their continuous guidance and support. I also thank Francesco D'Acunto, Thierry Foucault, Francesco Franzoni, Rachel Nam, and Alberto Rossi for insightful comments and discussions. The author declares no sources of financial relationships, or potential conflicts of interest related to this research.

1 Introduction

Asset managers are in the business of making predictions. It is not surprising then, that increasing availability of data and technological innovations are impacting the industry. For example, artificial intelligence or increasing computational power can affect the way managers select their assets. A growing literature studies how data abundancy and the adoption of new technologies impact managers' portfolio decisions. However, asset managers focus on more than just predicting returns. The majority of their compensation is tied to the total assets under management (AUM) and the inflows they can attract[1]. Therefore, a significant amount of their incentives is aligned toward increasing their total assets managed. These observations raise many interesting questions: Do asset managers employ development in new technologies to improve their ability to attract capital flows? How does this impact the structure of the asset management industry?

In this paper, I address these questions. Funds have increasing access to customers' information, and prediction models can help forecast future demand for products (e.g., thematic funds) or the maximum fees investors are willing to pay. For example, Broadridge, one of the leading FinTech companies in the US, offers a service called "*Distribution Insight*" as part of its asset management solutions. This service provides consulting and analytics on future demand for asset management products, information on which distribution channels are most effective for different types of funds (e.g., financial advisors or direct-to-consumer), prospect investors, and many analysis on both own and competitors' customers.

Most studies on the role of new technologies in asset management focus on their implications for the portfolio management side of the industry. Yet, fund managers have strong incentives to predict their investors' demand, and such innovations can improve their ability to do so. Modern "prediction machines" allow managers to better understand evolving investors' preferences and tailor their product offerings accordingly, ultimately helping them

[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.

attract more inflows. Little is known about this direct impact of new technologies on the asset management industry. While it is widely recognized that investors learn managers' skills over time, the implicit common assumption in the literature has been that no economic forces work in the opposite direction –treating asset managers' product menu as exogenous. This paper helps bridge that gap by showing that asset managers learn from investors' data and recent technological innovations have direct implications for fund managers.

A priori, the role of new technologies on asset managers' capital collection is not clear. The asset management industry faces search cost frictions (Hortaçsu and Syverson, 2004): investors spend time and other resources to gather information about managers' products. Costly search implies that investors buy their preferred fund within a subset of all the available products, as they are unaware of the entire menu of funds. This mechanism limits competition and generates price (fee) dispersion –even among homogeneous products. Theory predicts that as search costs fall (e.g., due to cheaper internet access), fee dispersion in the asset management industry disappears. If technological growth enables managers to lower these costs by offering clear, more timely information, competition will intensify, driving capital toward a more competitive market[2]. However, at the same time, new technologies may help fund managers better understand investor preferences, and cater to their specific demand. Suppose asset managers know more about what investors want. In that case, they might specialize or tailor product offerings, attract capital, and potentially charge higher fees –sustaining price dispersion. For instance, an asset manager could identify a growing demand for ESG products among young investors in specific geographic regions and use this knowledge to inform decisions about new fund launches or changing fee structures. Yet, whether this mechanism is at play is not obvious. Most models of rational markets for asset management do not leave space for managers to learn about investors' demand[3] (see Berk

[2]This is counterfactual in US data. Appendix Figure C.1 shows fee dispersion over time for Small Cap and Emerging Markets funds (Panel A, B) and the average dispersion *within* fund category for all US funds (Panel C). The dispersion in fees has been remarkably stable, if not increasing, since 2000.

[3]Some exceptions are Massa (2003) and Betermier et al. (2023). Loseto and Mainardi (2023) model a market in which individual mutual funds and fund families compete and can be interpreted as "managers' learning".

and Green, 2004; Gârleanu and Pedersen, 2018 among others). Thus, giving managers more information about investor preferences would have no impact whatsoever. Introducing an active role for investor preferences in product decisions suggests that new technologies may allow funds to avoid increasing competition by focusing on targeted products. Understanding these dual effects is important, as mutual funds and ETFs manage a significant share of household wealth (ICI, 2023) and are tightly linked to the efficiency of capital allocation.

I exploit information about asset managers' websites to quantify their willingness to collect and analyze investors' data. Every website is made by different building blocks, called *technologies*. For example, e-commerce often install technologies that allow safe payments, like Apple Pay or PayPal. My strategy is to identify technologies that collect and analyze customers' data (e.g., Google Analytics) and the exact installation date on asset managers' websites. These *data technologies* allow storing digitized information about user interactions over websites' pages, detailed demographics, and purchase history. They can help asset managers learn investors' product preferences or run A/B tests (a tool similar to Randomized Control Trials). I obtain data from BuiltWith, an alternative data provider specializing in website profiling. BuiltWith scans websites in search for clues that identify the usage of technologies, such as HTML tags in their code. They continuously crawl several websites and keep track of the installation date (and eventual removal) of a website's technologies. Thus, observing the adoption date of data technologies by asset managers, I have a proxy for when a given manager starts collecting investors' information. I test whether new technologies affect fund inflows for a sample of US equity mutual funds and ETFs from CRSP and Factset.

I find that asset managers receive more inflows after the adoption of data technologies. Funds using technologies that collect and analyze investors' information receive between 1.5% and 2% more flows each year. This effect is not negligible, as the unconditional average flow in my sample is indistinguishable from zero. The estimates are larger if I restrict the sample to only retail share classes. While I find no effect on institutional share classes. These results seem sensible since the data technology I study generate signals from web traffic data, which

are arguably more informative about retail investors. In addition, when asset managers collect investors' data, they charge higher fees, and the number of funds offered within the fund family grows.

As a first step, I provide evidence of asset managers attracting more capital after adopting data technologies. I use a difference-in-differences framework to study how asset managers change after installing technologies that collect customers' data. I compare changes in monthly inflows (pre- and post-adoption) within the same fund and month-category. This approach allows me to control for time-invariant fund characteristics and common shocks affecting funds or a specific style in a given month. First, I find that funds receive more flows after adopting a data technology on their website. Asset managers attract 0.14% larger monthly flows (around 1.5% yearly). This effect translates into +$40 million/year for the average US fund in my sample[4]. Furthermore, funds' marketing and distribution expenditures do not explain this effect[5].

Second, I find that funds charge higher expense ratios after installing data technologies. The effect is not driven by an increase in marketing and distribution fees. This result is important since it highlights that technological innovation does not always reduce search costs frictions[6]. Development in technologies such as smartphone apps might reduce investors' search costs, as they allow easier access to information at their fingertips. However, my results show that the impact of new technologies on the asset management industry is more complex than this. When *managers* learn about customer preferences and data, technological innovations allow funds to specialize and charge higher fees than their competitors. Furthermore, after adoption, there is no change in funds' performance (nor in value added Berk and van Binsbergen, 2015), suggesting that managers appropriate all the surplus.

Third, learning about investor preferences through web traffic data should enable fund

[4]This amounts to +$23 million/year in January 2000's USD.

[5]Results are robust to concerns over staggered difference-in-differences (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021).

[6]See, among many others, Thakor (2020); Basten and Ongena (2020); Hong et al. (2024); Argyle et al. (2022) for examples in which development in financial technologies reduce search costs.

managers to better predict demand and fill gaps in their product offerings. Consistent with this hypothesis, I find that at the fund family level, data technology adopters open more new funds than non-adopters in the period following adoption. Moreover, better prediction of customer demand indirectly affect portfolio decisions: adopters maintain a smaller cash buffer in their portfolio when expected returns are high, and hold more illiquid assets, suggesting that knowledge of future flows reduces redemption risk.

The above results do not warrant a causal interpretation, since the adoption of data technology is a *choice* made by asset managers. For example, even though I find no evidence of pre-treatment trends, managers might still install those technologies when they expect higher demand for their products. In this setting, asset managers' endogenous choice makes it challenging to identify a causal link. I address this concern using an aggregate shift in the information that funds can extract from web traffic data, which is plausibly exogenous to fund flows. Specifically, I exploit the public release of TensorFlow, a widely used open-source machine learning (ML) library that significantly increased the availability of ML worldwide[7]. Google's release of TensorFlow in November 2015 dramatically increased predictions' precision where large amounts of data are available. For example, Uber and Airbnb directly 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 for more informative predictions from data, the effect of data technology adoption, if any, should increase post TensorFlow release. I confirm this conjecture by estimating the additional effect of data technologies after the shock, removing from the analysis fund managers who adopt after TensorFlow's release. The impact of data is 30% larger after the release of TensorFlow, which plausibly allows extracting more precise signals from customers' data. Furthermore, I exploit cross-sectional heterogeneity *before* TensorFlow's release to confirm the intuition behind this result. I construct two proxies for funds' availability of data when TensorFlow is released: (i) the number of months between the fund's adoption of a technology and Novem-

[7]See, for example: wired.com/[...]/google-open-sources-its-artificial-intelligence-engine.

ber 2015, and (ii) the number of different data technologies installed. These two measures serve as proxy for the amount of data available, 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 the cross-sectional size of the dataset). Using a difference-in-differences specification with (i) and (ii) as continuous treatment intensity, I show that funds with larger datasets benefit more from TensorFlow's release. Importantly, these results do not assume that all funds must use ML algorithms to analyze their customers' data, as this setting is similar in spirit to an intent-to-treat (ITT) specification.

Finally, I run a battery of validation tests consistent with the hypothesis that my findings are driven by asset managers' extracting information from customers' interactions. First, one would expect that signals collected from web traffic are mostly informative about retail investors. I compare the effect of data technology adoption on retail share classes against institutional share classes, within the same fund. As expected, I find no effect on institutional share classes and a strong positive effect on retail share classes. Second, placebo tests on adoption of website technologies *not* aimed to collect and analyze customers' data (e.g., network technologies) give insignificant results. Third, I study the effect of competition. When many funds acquire similar information, the marginal benefit of it should decline. Accordingly, I show that the positive effect of data technologies decreases as more funds within the same category adopt those technologies. Fourth, consistent with concave returns to information (Veldkamp, 2011), the effects of adoption increase with the number of technologies installed, but the marginal benefit from installing an extra technology is decreasing.

I entertain 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, I formally test and reject this hypothesis using different measures of risk-adjusted performance. Another plausible explanation is that fund managers may not be learning about investor preferences, but they are *persuading* them. This mechanism would align with results being

purely driven by obfuscation and marketing (Mullainathan et al., 2008; Ellison and Ellison, 2009). If this were true, one key prediction would be that fund flows should be less responsive to changes in fees after adopting data technologies. Nevertheless, I reject this hypothesis, as fund flows are not less responsive to fee changes after technology adoption but rather more elastic.

Overall, my findings suggest that the impact of new technologies on the asset management industry is not limited to asset allocation. These results have important implications. For example, larger asset managers might disproportionately benefit from the bigger amount of data they can collect relative to competitors. On a similar note, this trend can affect managers' incentives. The efficiency of the asset management industry hinges on investors' ability to identify good funds and allocate capital to them. Using data, managers might reach their AUM capacity without fully aligning their incentives with investors'.

Related Literature. This paper contributes mainly to three strands of literature. First, it contributes to the growing literature on the role of new technologies in financial markets. Existing research studies the effect of these technologies in financial forecasting (e.g., Chi, Hwang and Zheng, 2024; Coleman et al., 2022; Dessaint, Foucault and Frésard (2024); van Binsbergen, Han and Lopez-Lira, 2022), stock market quality (e.g., Martin and Nagel, 2022; Farboodi and Veldkamp, 2020; Dugast and Foucault, 2024), households (e.g., Mihet, 2022; Rossi and Utkus, 2024; D'Acunto and Rossi, 2023) and capital allocation (Abis, 2022; Bonelli, 2024; Birru et al., 2024; Bonelli and Foucault, 2024). I study the effects of new technologies on asset managers' capital collection and their consequences on the industry structure. My results show that fund managers can use data technologies to gather and analyze investors' information and increase their inflows. In that sense, this work also relates to studies on the impact of technologies on the financial industry more broadly (Abis and Veldkamp, 2023).

Second, this paper contributes to the literature on the industrial organization of the asset management industry.[8] Hortaçsu and Syverson (2004) show that investors' search costs are

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

crucial to explain why homogeneous S&P500 index funds charge different fees. Intuitively, if investors face costs when searching for managers' products, their choice will not necessarily be to buy the best option available. They will pick the option within a subset of all available products, limiting competition and generating price dispersion even in homogeneous product markets. Roussanov, Ruan and Wei (2020) explore the role of mutual funds' marketing and distribution expenditures in a market with search costs. Their results highlight the importance of mutual funds' marketing for attracting investors' capital[9] (Sirri and Tufano, 1998; Reuter and Zitzewitz, 2006). According to this strand of literature, technological improvement will reduce search costs frictions and increase the industry's competition. More recently, Obizhaeva (2024) finds that ETFs attract more flows when advertising on online search engines. Within this literature, a few papers consider the strategic product market choices by asset managers (e.g., Massa, 2003; Betermier et al., 2023; Cvitanić and Hugonnier, 2022; Kostovetsky and Warner, 2020; Bonelli, Buyalskaya and Yao, 2024; Loseto and Mainardi, 2023). In these works, fund families change their product menus to reduce investors' switching costs or to differentiate themselves from others. To the best of my knowledge, this paper is the first to investigate whether asset managers collect and analyze customers' data to improve their ability to attract flows. These findings provide a framework for thinking about managers learning from investors' demand to optimize their product menu (either expanding the product space or changing fee structure).

Third, this paper contributes to the literature on the role of data in the economy (e.g., Jones and Tonetti, 2020; Cong et al., 2020; Brynjolfsson and McElheran, 2016; Goldfarb and Tucker, 2019). Chung and Veldkamp (2024) review this growing literature in detail. The main insight is that the increasing amount of digitized information is valuable for economic agents, and a no-data equilibrium is different from a data economy (Farboodi and Veldkamp, 2023). Although digitized information itself might not be different from simple information,

importance of the asset management industry for asset prices.

[9]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, Musto and Wermers (2014) for a survey.

what is different is the enormous amount of data points available and the sources from which agents extract those information. I show that investors' data are part of an asset manager's stock of knowledge and help funds cater to investors' demand. Fund managers can collect and analyze useful information about prospect investors from web traffic on their own websites. Therefore, valuable data for asset managers are not only datasets for identifying investment opportunities (Farboodi et al., 2021, 2024; Bonelli and Foucault, 2024), but customers' data can also represent a crucial part of managers' information set.

2 Economic Framework

To illustrate the main mechanism, I provide a simple and tractable framework of fund flows and fees in a market with heterogeneous investor preferences. The setup builds on a Hotelling-type of model, where investor preferences are described by a continuous variable. Without information on the distribution of customer preferences, two funds will equally share the flows to the asset management industry. However, when one fund uses data technologies to learn more about the distribution of preferences, it will offer products that are closer to what investors prefer. This mechanism allows the "DATA" fund to attract more capital and charge higher fees in equilibrium. The goal of this framework is to guide the empirical strategy and discuss additional predictions.

Model setup. There are two dates, $t = \{0, 1\}$. Two risk neutral asset managers, denoted by D and N, offer investment portfolios to a measure-one continuum of investors indexed by i who can only invest through the asset managers, i.e., investors do not invest directly in the asset market[10]. Here, D stays for DATA and represents the manager with data technology in place, while N is a naive fund without access to data technology. Without loss of generality, I assume each investor i picks exactly one asset manager. Importantly,

[10]This assumption follows, for example, Basak and Cuoco (1998) and Gârleanu and Pedersen (2018). A similar situation might arise because investors can share information costs or other types of frictions. One can think about the investors in this section as the fraction of investors that decide to invest via an asset manager, e.g., $\alpha = \mu/(\gamma\sigma_r^2)$ in a CARA-normal setup similar to Grossman and Stiglitz (1980).

investors have heterogeneous preferences. I summarize those preferences in a continuous variable represented on a line $x \in \mathbb{R}$. For example, x might represent a linear combination of several characteristics and tastes, the outcome of k-means analysis, or PCA, all commonly used methodologies in data and customer analytics (Abdulhafedh, 2021; Savic et al., 2019). In the asset management industry, a large x_i might represent investor i's preference for ESG stocks (Pástor, Stambaugh and Taylor, 2021) or specific themes such as AI- or cannabis-ETFs (Ben-David, Franzoni, Kim and Moussawi, 2022). At the same time, x_i might represent other personal tastes, such as the preferred distribution channel for buying a fund (e.g., through a financial advisor or brokerage account)[11]. I assume investors have quadratic costs in deviating from their preference x_i. Formally, the net cost investor i faces when buying fund j is:

$$c_{i,j} = f_j + t \cdot (x_j - x_i)^2, \quad \text{for } j = \{D, N\},$$

where f_j is the fee (price) charged by fund j, and t is a parameter that governs how costly it is for investors to deviate from their own preferences. A large t implies that investors perceive as expensive to buy a product much different from x_i. The term $(x_j - x_i)^2$ represents how investor i's taste is distant from fund j. Asset managers offer a variety of products to clients with heterogeneous needs, endowments, hedging motives, and preferences. Intuitively, investors will find buying a fund closer to their preference more attractive.

The mass of investors with preference equal to x_i is given by $\phi(\mu; x_i)$, where $\phi(\mu; \cdot)$ is the p.d.f. of a logistic distribution with average μ, and scale parameter one. Similarly, the c.d.f. of the same logistic distribution is denoted by $\Phi(\mu; \cdot)$.

Crucially, the average of the distribution, μ, is randomly drawn from a standard normal distribution: $\mu \sim \mathcal{N}\{0, 1\}$. If $\mu > 0$, the distribution of preferences $\phi(\mu; x)$ shifts rightwards, which means that a larger mass of investors prefers products with $x > 0$. Investors know their preferences, while fund managers only know the distribution (i.e., they don't know the

[11] Industry practitioners in the retail funds segment are particularly interested in understanding the best distribution channel for a given clientele, see for instance: broadridge.com/press-release/2025/.

realization of μ). Thus, they don't observe the actual profile of customers' tastes. However, the asset managers using data technologies, D, extract a signal s from collecting and analyzing investors' data in $t = 0$. The signal is unbiased and noisy[12]: $s|\mu \sim \mathcal{N}\{\mu, \sigma_s^2\}$; I will denote a signal's precision as $\tau_s = \sigma_s^{-2}$. Collecting data allows a manager to learn investors' tastes better, and the lower σ_s^2, the more data are informative. The unknown realization of μ determines where most investor preferences are, and having a signal on it makes data technologies relevant in this simple framework. Figure 1 shows an overview of this simple framework for $\mu > 0$. In what follows, I discuss the case of $\mu > 0$: in this case, a larger mass of investors will lie on $x > 0$. The same economic intuition will hold for $\mu < 0$, with a slightly different formulation.

In $t = 0$, the two asset managers D and N set up a fund picking one location over the line of preferences denoted x_D and x_N, respectively, to align as much as possible with investor demand. I assume the costs of setting up a fund are independent of x, so both funds have the same marginal cost c, regardless of their location choice on the line. In $t = 1$, funds compete on fees to maximize profits, and each investor i buys the fund that is most convenient for her.

2.1 Summary of the Equilibrium

The fundamental intuition in this simple example is that funds can differentiate their product offerings to better meet investors' demand. Asset managers located closer to a larger mass of customers will receive more capital. An equilibrium in this framework is a set of fund locations (x_D, x_N) and fees (f_D, f_N) such that both funds pick the closer x to their expected μ in $t = 0$ since they minimize the distance to the mode of investor tastes, and maximize their respective profits in $t = 1$. Importantly, I separate the location and pricing problems, as it is well known that Hotelling-type of models have no equilibrium when firms can use both location and prices (contemporaneously) as strategies, see for example (Shy,

[12]This simple example does not include endogenous information choice; its only purpose is to illustrate the idea behind the hypothesis that managers can learn valuable information from customers' data.

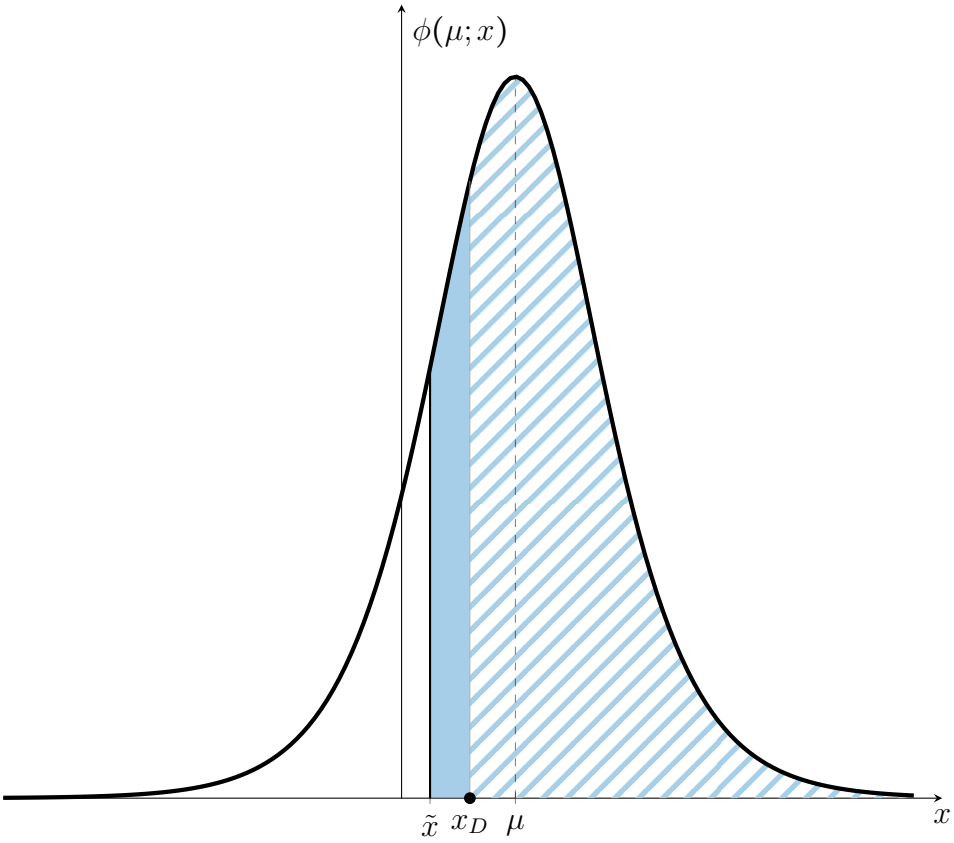


FIGURE 1: **Economic Framework Overview**. This figure illustrates an overview of the distribution of investor preferences $x \in \mathbb{R}$. The black curve shows $\phi(\mu; x)$ for $\mu = 2$. x_D is the location picked by the "DATA" fund D, and $\tilde{x}$ the threshold below which investors will find N more convenient (see Section 2.1). The area underneath the p.d.f. $\phi(\mu; x)$, on the right of $\tilde{x}$, represents the total market share of fund D (filled area). The hatched area shows fund's D captive market share, whereas the shaded lightblue area (vertical section between $\tilde{x}$ and x_D) is the fraction of market share that D gains competing with N on fees.

1995, Proposition 7.7). For instance, this simplification might imply that once funds commit to delivering a particular investment strategy, they find it costly to deviate from it. Abis and Lines (2024) find consistent evidence in the US mutual funds industry. In what follows, I will first consider a baseline equilibrium in which none of the two funds can observe a signal about the distribution of investor preferences. This benchmark is useful for comparison with the second equilibrium, in which I allow the "DATA" fund D to extract a signal from collecting and analyzing customers' data. Appendix B contains all the proofs.

Baseline. First, consider a benchmark economy where no data technologies are available. In this case, asset managers D and N know only the prior distribution of μ (i.e., as if

$\tau_s \to 0$). Thus, in $t = 0$, when funds set their location on the preferences line, they will choose $x_D = x_N = \mathbb{E}[\mu] = 0$. Intuitively, as fund managers have no information about where the realization of investor preferences is, they will locate on their point where they believe more investors lie –i.e., on the mode (and average) of the logistic distribution with $\mu = 0$. In $t = 1$, with both funds in the same location $x = 0$, competition will resemble a Bertrand duopoly equilibrium. The two funds will split the total share of investors' flows equally and price at marginal cost c.

Next, I consider a similar market for asset management products, where fund D observes a signal on the realization of investors' preferences.

Asset and Data Managers. When fund D has access to data technologies, the manager receives a signal s about the realization of investor preferences. Now, while in $t = 0$ the naive fund N will pick location $x_N = 0$, the "DATA" fund will set $x_D = \frac{\tau_s}{1+\tau_s} s$. Thus, when observing the unbiased signal, asset manager D deviates from $x = 0$ to locate closer to the mode of investor preferences. The deviation in x can be interpreted as specialization vis-à-vis a "standard" product. In the asset management industry, funds can differentiate from competitors in many ways; for instance, they can overweight stocks with particular characteristics, change their distribution channel, or differentiate their fund prospectus.

Once D sets $x_D > 0$, the mass of investors on the right-hand side of x_D will find the fund offered by D more attractive, as it is closer to their preferences. This market share is given by $1 - \Phi(\mu; x_D)$, and it represents fund D's captive investors (hatched area in Figure 1). Similarly, investors on the left-hand side of $x_N = 0$ will prefer fund N. This mass of investors is determined by $\Phi(\mu; 0)$, representing fund captive demand for fund N. However, the mass of investors between x_N and x_D is not captive for either of the two funds. In this area, there will be a threshold $\tilde{x}$ above which investors find D more attractive than N, and below, they will prefer N. This threshold is given by:

$$\tilde{x} = \frac{x_D}{2} + \frac{f_D - c}{2x_D t}. \tag{1}$$

Intuitively, $x_D/2$ represents the midpoint between the products offered by funds D and N. Then, the threshold $\tilde{x}$ shifts to the right –resulting in lower market share for fund D– as the "DATA" fund increases its fee over its costs. Investors in this area prefer fund N if f_D is too large, even though their preferences might be closer to x_D.

In choosing fund fees f_D, the "DATA" fund manager maximizes its total profits, $\pi_D = [1 - \Phi(\mu; \tilde{x})] \cdot (f_d - c)$, trading-off a higher market share from setting a lower f_D, with fewer profits. The optimal fee satisfies:

$$f_D = c + 2x_D \cdot t \cdot \exp\{-(x_D - \mu)\}. \tag{2}$$

A key implication from equation (2) is that fund D can set its fee above marginal costs. The second term in the equation determines how much the "DATA" fund charges in excess of fund N (without data technologies), in equilibrium. This term increases with the fund's deviation from $x = 0$, and it is larger as x_D lies closer to μ (i.e. when the signal from data is more precise and the fund D locates closer to a larger mass of investors).

2.2 Discussion and Testable Implications

This simple example yields intuition for guiding the empirical specification. To summarize, the framework has two main predictions.

Hypothesis 1. (Flows and Data Technologies) *Asset managers with more information about investor preferences receive a larger share of the total flows to the asset management industry.*

In the example above, the market share is $(1 - \Phi(\mu; \tilde{x})$. Since $\mu > 0$, $\Phi(\mu; \tilde{x}) < 0.5$, and the fund employing data technologies collects a larger share of the total flows accruing to the asset management industry. As funds with data technologies in place offer products closer to a larger mass of customer preferences, they will collect more capital. When bringing this framework to the empirics, I measure the asset management "market" as the total inflows

to the industry. Therefore, my first hypothesis posits that funds receive higher flows after adopting a data technology (i.e., after collecting signals on investor preferences).

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

When a fund's product offering is closer to a larger mass of investor preferences, those investors are willing to pay more for the product, and the manager can charge a higher fee. Moreover, in a setting where the fee set by funds without data technologies $f_N = c$ is equal to investors search costs (e.g., as in Gârleanu and Pedersen, 2018), the difference $f_D - f_N$ is equivalent to the dispersion in fund fees. Thus, the dispersion in fees charged by funds in equilibrium can remain large even though search cost frictions decline. This simple example provides intuition consistent with what we have observed empirically in the past years. Even though search cost frictions have arguably been declining (e.g., because of cheaper internet access), the dispersion in fund fees within a given category did not decrease at all. Figure C.1 in the Appendix shows the fee dispersion within fund categories from 1995 to 2022. In this modern sample period, the dispersion in fund fees has been remarkably stable, which is at odds with theories purely based on search cost frictions. This mechanism is consistent with results in Bonelli et al. (2024) and reminiscent of Menzio (2023). In particular, Menzio (2023) develop a model where firms produce more specialized products in an environment with declining search costs. Selling more specialized products allows firms to charge higher prices in equilibrium and maintain price dispersion along a balanced growth path.

While this example abstracts from some important considerations, it illustrates *how* data technology can help funds to attract capital. Asset managers with more information about investors' tastes can collect more flows $(1 - \Phi(\tilde{x})$ and charge higher fees at the same time.

Fees, Redemptions, and Number of Funds. The availability of data technologies to inform managers on investor preferences have a few additional testable implications for the asset management industry. First, if a data technology informs managers about their investors' preferences, it may enable to maintain a lower cash buffer during bad times (when

expected returns are higher but redemption volatility is large). Intuitively, funds with better knowledge of their customers have lower uncertainty about fund liquidation. Thus, when more investment opportunities are available (in states of the world with higher expected returns) managers may face less redemption uncertainty and maintain a lower cash buffer in their portfolios. On a similar note, funds might hold more illiquid assets, as it happens with more "committed" capital (Gómez, Prado and Zambrana, 2024).

Second, the information advantage gained through investors' data may influence product development strategies. For instance, suppose funds predict a surge in demand for thematic products. In that case, they might be better positioned to timely offer those funds to investors[13].

3 Data and Measurement

In this section, I describe the data and report a series of new facts on asset managers' adoption of data technologies.

3.1 Data

I focus on US mutual funds and ETFs from March 1993 to December 2022. My main data sources are the CRSP Survivorship Bias-Free Mutual Funds dataset and Factset, to identify share classes of the same fund. I also use Thomson Reuters[14] (s12) holdings, and information from N-SAR filings reported at the SEC. I merge CRSP/Factset by *cusip*, which identifies unique financial securities and is not re-assigned over time. I follow Berk and van Binsbergen

[13] Ben-David et al. (2022) rationalize the increase in offerings of thematic products as "competition for attention". In their paper, ETFs compete on either price (fees) or attention (specializing in the product space). While their findings align with this paper's results, I provide another perspective on the proliferation of thematic funds.

[14] 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 (Appendix Figure C.2). Thus, I prefer not to append different data sources and use the updated s12 holdings throughout the sample.

(2015) as closely as possible in cleaning CRSP funds' data. I summarize here the main steps and provide thorough details in Appendix A. I select all US equity mutual funds and ETFs. I adjust all AUM numbers by inflation (in January 2000 dollars). I remove observations dated before the fund's first offer date to reduce incubation bias concerns (Evans, 2010). I also drop funds with less than two years in the sample and before their total (inflation-adjusted) AUM reaches $5 million for the first time (Berk and van Binsbergen, 2015). CRSP 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, summing up the AUM of all subclasses and weighting all other variables (e.g., fees, returns) by lagged AUM. The sample start date (March 1993) is dictated by the limited availability of AUM before that date (see Pástor, Stambaugh and Taylor, 2015). The final sample includes 7,900 funds (7,649 equity mutual funds and 251 ETFs) and 987,242 fund-month observations. My sample is somewhat larger than comparable samples using CRSP mutual funds data. The reason is that I do not remove index funds, institutional share classes, sector funds, or funds that allocate less than 80% of their portfolio to stocks. I estimate funds' alphas using rolling-window regressions from monthly returns in the past 24 months. The holdings' sample starts in 2004:Q2, because the SEC began requiring quarterly holdings disclosure in May 2004, following the adoption of Rule 30b1-5.

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

$$Flow_{i,t} = \frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t}) - MRG_{i,t}}{AUM_{i,t-1}} \times 100, \tag{3}$$

where $AUM_{i,t}$ is 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 to misattribute funds' mergers as inflows. Finally, I winsorize all variables at the 1% and 99% level.

Table 1 about here

Table 1 presents summary statistics for all fund-months observations in my sample. The distribution of AUM is rightly skewed, as is common in the institutional investors' literature. The average expense ratio is 1.13%, with 0.28% attributable to marketing and distribution expenses (12b-1 fees). On average, funds in my sample have statistically insignificant 0.12% outflows each month and net abnormal returns (after expenses) are zero or negative on average, consistent with evidence for the US.

3.2 Data Technologies

To identify asset managers' willingness to collect data, I exploit information contained on their websites. Specifically, I use the website's code to detect whether asset managers adopt technologies for storing and analyzing investors' web traffic data. Web traffic data are widely used in several industries to understand customers' behaviors and preferences. Therefore, this is a good laboratory for studying the role of data technologies in asset management.

Every website is made by different tools that work as building blocks. These building blocks are typically called *technologies*. For example, installing Google Maps technology allows a website to display an interactive map on its pages (e.g., to show store locations). Other technologies, like Adobe Analytics, are aimed to collect and analyze web visitors' data. I obtain technologies' adoption data from BuiltWith, an alternative data provider, and use the adoption date of analytics technologies as a proxy for managers' willingness to collect customers' data. BuiltWith analyzes websites' source code and searches for specific patterns, such as HTML tags, that identify the presence of technologies. They continuously crawl websites to capture installed technologies, starting January 2000. Therefore, I observe the exact month a website installs (and eventually removes) a technology. Henceforth, I define *data technologies* as those aimed to collect and analyze visitors' data, such as Google Analytics or Adobe Analytics. These tools are specifically designed to gather or generate signals from users' data. My goal is to study whether the ability to collect capital differs between asset managers adopting data technologies and non-adopters.

I merge BuiltWith data with all funds' websites from CRSP. Moreover, since CRSP reports funds' websites starting January 2008, I hand-collected information on the website registration date from `whois.com`, and back-fill a fund's website until its registration date[15].

Figure 2 and Table 2 about here

Figure 2 shows the adoption of data technologies in my sample of funds. The blue area represents the total number of funds with at least one data technology on their website each month. The red line shows the percentage of funds adopting data technologies. A few early adopters started employing these technologies since 2006-2007. However, the big shift in data technology adoption occurred in 2012, when the percentage of funds with such technologies grew from 15% to 30% in the cross-section. This increase is concurrent with a major release of Google Analytics by Google, which is still the leading data technology provider nowadays. Data technologies' adoption in asset management surged significantly after 2012 and gradually stabilized after 2017. As of December 2022, as many as 75% of asset managers in my sample (more than 2,000 unique funds) have adopted technologies to collect visitors' web traffic data.

In Table 2 I report the leading data technologies by adoption, as of end-of-sample. Google Analytics accounts for the lion's share of adoption, with around 60% of funds installing it[16]. Among other leading technologies, I find Omniture Test & Target, used for A/B testing[17], and LiveRamp, which helps the storage of big data. Overall, all technologies in Table 2 are used to collect or generate informative signals about web visitors' preferences. This features are not limited to the most common data technologies used by asset managers in my sample. Appendix Figure C.3 shows the word cloud built from descriptions of all data technologies installed by asset managers.

[15]This procedure marginally increases the number of technologies' adoptions in my sample (see Figure 2). However, it ensures I don't misclassify funds as non-adopters before January 2008.

[16]Appendix C.2 ensures the results are not entirely driven by Google Analytics (see Tables C.6 and C.7).

[17]A/B tests are tools similar to RCTs, which are increasingly used in several industries. They allow to randomly split web traffic audience and study several alternatives of products' bundles, pricing, etc.

Technology Diffusion in the Asset Management Industry. The diffusion of technological innovation is critical for understanding growth and improvements in several industries (Barro and Sala-I-Martin, 1997). The dynamic of technology adoption in the asset management industry is particularly important, as it might inform us about the dynamics of information diffusion. For example, technology diffusion (and information) may spread among peers within a particular fund style (e.g., among competitors) or geographically (e.g., within a given state or city through social interactions). Given that precise data on technology adoption are typically not readily available (Stokey, 2020), I briefly examine how data technology diffuses among asset managers. I explore whether adoption is driven primarily by geographic location or fund category by regressing the probability of installing a data technology, on the (lagged) adoption rate within a given geographic cluster or fund category. In particular, I define the adoption rate in a fund category for month t as the percentage of funds within that category with data technology in place as of month t. Similarly, I define adoption rate within US states, cities, and zip codes. Then, I regress the probability to adopt data technology on the lagged adoption rate within fund category and geographic location, using a probit/logit specification[18]. The results are in Appendix Table C.8. Interestingly, the adoption rate within fund style is never significant in predicting the probability of adopting data technology in the subsequent month. On the other hand, adoption rates at the state, city, and zip code levels consistently show a positive relationship with the probability of installing data technology. This result is in line with previous research on the information diffusion among asset managers (Christoffersen and Sarkissian, 2009; Cujean, 2020). While it is important to emphasize that technological adoption differs fundamentally from the diffusion of *ideas* and information, these results suggest a potential relationship between technology and information diffusion in asset management. Clearly, this evidence is only suggestive and by far not conclusive. Exploring how technology adoption spread in the asset management industry remains a relevant area for future research.

[18]To avoid double counting, I exclude all observations after the first adoption month from the logit/probit regression. Given the persistence in technology use after initial adoption, this correction is important.

4 Empirical Findings

In this section, I use information on asset managers' adoption of data technologies to explore the impact of new technologies on the industry. Guided by the economic framework in Section 2, I first test the effect of data technology adoption on fund flows and fees. Then, I examine ancillary results consistent with fund managers extracting useful information from customers' data.

4.1 Funds Flows and Data Technologies

The first main hypothesis from the framework above (Hypothesis 1, Section 2.2), argues that if data technology allows an asset manager to collect signals on investors' preferences, it will ultimately be reflected in the fund's inflows. Therefore, I first study whether funds receive more capital after adopting a technology analyzing web traffic data. 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 in that month. Then, I estimate the following baseline specification:

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta \ \text{DATA}_{i,t} + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \tag{4}$$

where $Flow_{i,t+1}$ is the fund i's flow in month $t+1$, defined in equation (3). α_i and η_t are fund and time fixed effects, respectively, and $\boldsymbol{X}_{i,t}$ is a set of control variables. The control variable vector contains fund size ($logAUM$), performance, (log) age, expense ratio, turnover, and monthly flows. I proxy for fund performance with the yearly CAPM alpha, as it is the closest model to the one investors use to make capital allocation decisions (Berk and van Binsbergen, 2016). Results are similar using other measures such as Fama-French 3- or 5-factors alpha (see Appendix C.2). The main coefficient of interest is θ. Importantly, including fund and time fixed effects implies that identification of θ comes from variation in flows before versus after data-adoption, relative to the same change for other funds without data technology. Standard errors are clustered at fund and month-levels.

Table 3 about here

Table 3 reports the main results from regression (4). I omit coefficients on control variables for brevity[19]. The first row shows the coefficient on the dummy variable $\text{DATA}_{i,t}$. The baseline result of 0.14 (t-stat 3.39) in column (1) suggests that funds adopting technologies aimed to store and analyze web traffic data are associated with larger fund flows. The coefficient implies a 0.14% (1.68%) larger monthly (yearly) inflows for data-driven funds after adoption. In columns (2) and (4), I include category×time fixed effects to account for any category-specific shift in a given month. The coefficients are consistent with the baseline specification in equation (4) (t-stat 3.14), suggesting that results are not driven by one category. I further address the concern that results might be driven by funds marketing and sales expenditures. If adoption of data technologies is associated with a fund's re-branding or increasing marketing efforts, my findings might actually reflect the effect of marketing in attracting fund flows (Sirri and Tufano, 1998; Roussanov, Ruan and Wei, 2020). To rule out this alternative explanation, in columns (3) and (4), I include 12b-1 $\text{Fees}_{i,t}$ in the set of covariates, which represent marketing and distribution expenses incurred by the fund. As expected, marketing and distribution fees are positively related to fund flows (second row). However, the effect of marketing on flows does not explain the coefficient on data technologies (t-stat 3.47). Therefore, funds adopting data technologies experience larger fund flows above and beyond the role played by marketing efforts[20].

Moreover, as the adoption of data technologies is staggered over time, in Table C.1 I verify these results are robust to recent critiques on staggered difference-in-differences (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021). The results are qualitatively and quantitatively unchanged[21]. Notably, the magnitude of the

[19] All controls enter with the expected sign across all specifications, e.g., positive flow performance sensitivity (Sirri and Tufano, 1998; Pástor et al., 2015; Franzoni and Schmalz, 2017).

[20] In Section 5.1, I provide further evidence that my results are unlikely to be explained by re-branding or marketing.

[21] In my setting, the total weight attached to "forbidden comparisons" in staggered diff-in-diff (Goodman-Bacon, 2021) is only 8.1%. This is because my sample starts several years before the first adoption of data

estimated coefficient for θ and its t-stat remains nearly identical across all specifications.

The fund flows described in equation (3) represent *net* flows, which capture the difference between inflows and outflows for fund i during month t. To further investigate the impact of data technology on fund flows, I exploit information from SEC regulatory filings (NSAR), which require funds to report their monthly inflows and outflows separately. The NSAR sample runs from January 2006 to June 2018, at which point NCEN filings replaced NSAR. For additional details on these SEC regulatory filings, see Evans et al. (2024). Appendix Table C.2 shows the results for fund inflows and outflows, separately. In line with my Hypothesis 1, funds experience larger (new) inflows following the adoption of data technologies (columns (1) to (3)), while there is no significant difference in the funds' outflows post-adoption.

To study the dynamic of the effect of data technologies on flows, I interact the coefficient of interest in Table 3, with event-time dummies for each month before and after adoption. Figure 3 shows the estimated coefficients with 95% confidence intervals. Importantly, there is no evidence of significant pre-trend before installing a technology. Moreover, the effect of data technology on fund flows is persistently significant after 8 months from adoption.

Figure 3 about here

In Appendix C.2, I run tests for the parallel trends assumption, implicitly behind the identification in equation (4), and balance covariates. I find no evidence of statistically different pre-trends (Table C.3) or imbalanced covariates before adoption (Table C.4).

Taken together, these results support the view that data technologies help asset managers attract more capital. Funds adopting data technology are associated with larger inflows after adoption, and this result goes beyond the role of marketing expenditures.

technology (treatment), which allows precise estimates of group effects (Gardner et al., 2024). A similar argument applies for period (time) effects.

4.2 Increase in Expense Ratio

My second hypothesis in Section 2.2, is that fund managers with more data about customer preferences should charge higher expense ratio, all else equal. For example, if these technologies allow managers to offer funds closer to investors' personal preferences, they might elicit a higher willingness-to-pay and increase fees.

Interestingly, theory of the industrial organization of asset management predicts that equilibrium fees are determined by investors' search costs (see Hortaçsu and Syverson (2004); Gârleanu and Pedersen (2018); Roussanov et al. (2020), among others). A key prediction in these models is that when investors' search costs decrease, fees should decrease too. Intuitively, with lower search costs, investors can contact more managers, identify good ones, and the economy approaches the first-best with no price dispersion and equilibrium fees equal to the fund's marginal cost. According to this prediction, new technologies reduce fees for financial services (e.g., FinTech). However, the data technology I study in this paper differs from traditional financial technology. Data technologies do not facilitate information acquisition for customers but for asset managers. Thus, the prediction on funds' fees is the opposite of traditional financial innovation studied in the literature. When fund managers better align their product offerings with investors' demand, they can elicit a greater willingness-to-pay for products closer to personal preferences.

Table 4 about here

I study this hypothesis with a similar specification to equation (4), where the expense ratio (in %) is on the LHS. Table 4 shows the results. The first row reports coefficients on the dummy variable $\textsc{Data}_{i,t}$. As predicted, funds installing data technologies are associated with higher fees after adoption. The estimates in column (1) imply that funds using a data technology increase their fees by 1.5 basis points after adoption. Results are similar, including category×time fixed effects and controlling for marketing fees. Therefore, the increase in expense ratio is independent of changes in marketing and distribution expenditures.

Overall, these results are consistent with fund managers charging higher fees for funds closer to customers' preferences. Below, I explore other predictions consistent with the hypothesis that managers learn from customers' information.

4.3 Ancillary Results: Cash Buffer, and Number of Funds

In this section, I study further how new technologies can impact the asset management industry. First, managers better informed about customers' demand may face less redemption uncertainty. Therefore, I explore whether data technologies allow maintaining a smaller cash buffer when investment opportunities are available. Second, I study whether the product menu offered by the fund family expands to cater to specific preferences and diversify with respect to peers.

Cash Buffer. If data technologies allow for extracting informative signals on customers' demand, installing such technologies should help funds better predict future demand shocks. Therefore, adopting funds should face lower redemption uncertainty. To test this hypothesis, I study whether (after adoption) funds maintain a smaller cash buffer in their portfolio when expected returns are high. The intuition is that when good investment opportunities are available, an asset manager with no uncertainty about investors' redemption will try to take advantage of those opportunities and keep only a small part of her AUM to accommodate unexpected liquidations. On the other hand, if the manager has no information on investors' redemption, she might decide to maintain a larger cash buffer –forgoing investment opportunities. I study this prediction using the dividend-price ratio (D/P) to proxy for investment opportunities. Specifically, I run the following regression:

$$w_{i,t}(cash) = \alpha_i + \lambda_1 \cdot D/P_t + \lambda_2 \cdot (D/P_t \times \text{DATA}_{i,t}) + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t}, \tag{5}$$

where $w_{i,t}(cash)$ is fund i's portfolio weight in cash (in %), at month t, and $\boldsymbol{X}_{i,t}$ is the same fund-month controls' vector as in my main results. Intuitively, when D/P_t is high,

equity prices are (relatively) low and expected returns are high. Therefore, I expect λ_1 to be negative as asset managers might reduce their cash holdings when good investment opportunities are available. The main coefficient of interest is on the interaction term (λ_2). Importantly, identification of λ_2 comes from variation over time within a fund, not from variation across funds.

Table 5 about here

Table 5 shows the results. The coefficient estimates for λ_1 and λ_2 are in the first and second rows, respectively. Column (1) confirms that when the dividend-price ratio is high (higher expected returns), fund managers tend to reduce their cash holdings –although with only 10% statistical significance. Columns (2) and (3) include the main coefficient of interest: the interaction term 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 view that data technologies allow to extract useful signals about investor preferences. Asset managers with better precision about customers' demand face lower redemption uncertainty. Consequently, they can maintain a smaller cash buffer and deploy capital in investment opportunities.

On a similar note, data technology adoption is associated with funds tilting their holdings towards illiquid stocks, as measured by the Amihud illiquidity ratio (Appendix Table C.10). This result is in line with previous research showing that funds with more "committed" capital hold more illiquid stocks (Gómez, Prado and Zambrana, 2024).

Number of Products. Next, I test whether fund families increase their product offerings when funds within the family adopt a data technology. To test this prediction, I aggregate observations within the same fund family and test whether the number of funds within the family increases after adopting a data technology. I estimate the following specification[22]:

$$log(\text{N. of Funds}_{f,t+1}) = \alpha_f + \eta_t + \delta\ \text{DATA}_{f,t} + log(\text{Age}_{f,t}) + \varepsilon_{f,t+1}, \tag{6}$$

[22]Importantly, as the argument of the *log* in my specification is never zero, it does not suffer for well known identification challenges when the argument can equal zero (see Chen and Roth, 2023).

where N. of Funds$_{f,t+1}$ denoted the number of funds offered by fund family f in month $t+1$. Similarly to previous specifications, I define a dummy variable DATA$_{f,t}$ which takes value one if at least one fund within family f has a data technology in place in month t. I control for the fund family's age as it is plausible that families develop organizational skills over time, which reduces the cost of setting up a new fund. As for the results above, including family and time fixed effects ensures 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. I report results in Table 6, column (1).

Table 6 about here

Table 6 confirms that fund families with at least one data technology installed increase their product menu after adoption. More precisely, the semi-elasticity coefficient suggests that fund families offer 8% more funds after adoption. In column (2), I estimate a similar specification with a Poisson regression where the LHS is N. of Funds$_{f,t+1}$ (not in *log*). Results are similar, although statistical significance is reduced to 10%.

This result is consistent with the effect of other new technologies, such as AI, on firms' product portfolio. Babina et al. (2024) finds that more AI-intensive firms expand their product varieties, as AI facilitates the accumulation of knowledge and reduces uncertainty in product innovation.

Overall, these findings support the view that the role of new technologies in asset management extends beyond portfolio allocation decisions. The results in this section are consistent with fund managers adopting technologies to learn useful signals from investors' web traffic behavior, which in turn helps their capital collection efforts. Moreover, asset managers adopting technologies charge higher fees, keep smaller cash buffer when investment opportunities are available, and increase the number of funds offered within fund family.

4.4 Identification: Open Source Machine Learning

The above results suggest a positive relationship between data technologies and fund inflows, but do not establish causality. Adoption is an endogenous choice made by fund managers, so it might be correlated with other fund characteristics that cause fund flows. In this section, I exploit variation in the information that funds can extract from web traffic data, which is plausibly exogenous to investors' flows. Specifically, I use the public release of TensorFlow in November 2015, a major open-source machine learning (ML) library. The release of TensorFlow drastically decreased the cost of training and using ML algorithms in settings with large amounts of data available. Intuitively, ML allows fund managers to extract more informative signals from a given dataset. Accordingly, the effect of data technology on fund flows should increase after TensorFlow's release in November 2015.

The goal of my identification strategy is to isolate plausibly exogenous shock in the precision of information that managers can leverage to learn about their investors' preferences. Therefore, I compare the effect of data technology before and after TensorFlow's release, within fund. To alleviate endogeneity concerns about the adoption of data technologies after the shock, I remove from this analysis: (i) all funds installing data technology after November 2015, and (ii) funds installing technologies in the six-months window before TensorFlow's release[23]. This filters ensure that results are not driven by asset managers choosing to adopt a data technology, anticipating the release of TensorFlow.

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

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 \ \text{DATA}_{i,t} + \theta_2 \ \left(\text{DATA}_{i,t} \times \text{P}ost_t\right) + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \tag{7}$$

where the coefficient θ_2 captures the additional impact of installing a data technology, after

[23] Results are unchanged extending this window to 12 months.

TensorFlow's release.

Next, I exploit cross-sectional heterogeneity in funds to define a continuous treatment. Ideally, I would construct a continuous treatment based on the amount of investors' data available to each fund (just before TensorFlow's release). The intuition is that asset managers with larger datasets can benefit more from ML algorithms, and therefore, one should observe a more significant incremental effect for those funds. Since I cannot directly observe the amount of data available as of TensorFow's release date, I construct two fund-specific measures to proxy for it. The first proxy is the tenure of data technology adoption for each fund i as of November 2015. I construct a continuous variable equal to the (log) number of months between fund i's first adoption of data technology and TensorFlow's release date. Intuitively, this proxy captures the length of the time series data that asset manager i can use to train ML algorithms. The second proxy is the number of different technologies installed as of November 2015. The idea behind this proxy is that funds with more data technologies can collect more customers' characteristics and, thus, have larger datasets. Then, I study the impact of TensorFlow's release on data technologies using the following specification:

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 \ \text{DATA}_{i,t} + \theta_3 \ \left(\text{DATA}_{i,t} \times Post_t \times z_i\right) + \theta_4 z_i + \boldsymbol{\beta}' \boldsymbol{X}_{i,t} + \varepsilon_{i,t+1}, \qquad (8)$$

where z_i is either the tenure of data technology 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. Table 7 reports the results.

Table 7 about here

The second row shows results for θ_2 in specification (7). Coefficients estimates on this interaction term in columns (1) and (2) show that the effect of data technologies on fund inflows is 30% higher after TensorFlow's release. The third row reports estimates for θ_3. Columns (3)-(4) and (5)-(6) show the results using as continuous treatment the tenure of technology adoption, and the number of technologies installed, respectively. Results using

both continuous treatment confirm the intuition that the additional effect post-November 2015 is higher for funds with larger datasets.

A limitation in my identification is that by design, it relies on plausibly exogenous variation in prediction precision, not variation in data availability. With this caveat in mind, I think it is plausible in this context that more precise predictions are economically akin to more data/signals (i.e., ML increases the signal-to-noise ratio in customers' data).

Taken together, these results mitigate endogeneity concerns linked to the adoption choice, and they are consistent with a causal interpretation of data technology on asset managers' ability to attract capital.

5 Learning Mechanism: Validation and Robustness

The results above corroborate the hypothesis that asset managers use data technologies to learn investors' preferences and better attract flows. To further support that these results stem from a learning mechanism, I investigate other robustness tests to validate my hypothesis. First, I study the connection between technology adoption for retail and institutional share classes separately. Since my hypothesis hinges on asset managers observing signals from web traffic data, they should generate more information about retail investors. Second, I run placebo tests on different technologies that are not aimed at collecting and analyzing data. Third, I explore whether the returns to data technology are concave. Fourth, I study whether managers' competition in data collection impacts their ability to attract more flows.

Retail and Institutional Share Classes. The type of data technologies I study in this paper are designed to collect and analyze information from web traffic data. Accordingly, if fund managers improve capital collection by observing signals on customers' demand from these technologies, one would expect a more significant effect on retail investors. Intuitively, web traffic data should reveal more information about retail investors than institutional investors. Mutual funds data offer a good laboratory to test this prediction, as I can compare

the effect of data technology on retail and institutional share classes *within fund*. Specifically, I run a specification similar to equation (4), but at the share class level:

$$Flow_{j,i,t+1} = \alpha_i + \eta_t + \theta \text{ DATA}_{j,i,t} + \theta_R \left(\text{DATA}_{j,i,t} \times \text{Retail}_{j,i,t}\right) + \boldsymbol{\beta}' \boldsymbol{X}_{j,i,t} + \varepsilon_{j,i,t+1}, \tag{9}$$

where $Flow_{j,i,t+1}$ is the flow of share class j, in fund i at month t. I separate retail and institutional share classes and denote by $\text{Retail}_{j,i,t}$ a dummy equal to one if the share class j of fund i is sold to retail investors. Importantly, including fund fixed effects allows me to compare the effect on different share classes within the same fund. The coefficient of interest is θ_R, which captures the coefficient of data technology on retail share classes. If managers extract valuable signals from web traffic data, I expect θ_R to be positive. Table 8 shows the results.

Table 8 about here

The first row in Table 8 confirms that the effect of data technologies I study in this paper is concentrated on retail investors. The coefficient θ^R is positive and significant across all specifications. Moreover, data technologies have no effect on institutional share classes (second row). This result is consistent with websites' technologies being useful to generate signals about retail investors, but not enough to cater to institutional investors' preferences[24]. These results support the view that asset managers can learn customers' preferences to attract capital inflows.

Placebo tests. Next, to strengthen the interpretation of the results, I run a falsification test on my main result using technologies *not* aimed to collect and analyze data. That is, I estimate regression (4) substituting $\text{DATA}_{i,t}$ with placebo technologies installed on funds websites[25]. These technologies include, for example, Feeds (e.g., used to publish blog pages),

[24]Results in Table 8 are unchanged when aggregating retail (institutional) share classes separately, and running the regression at the fund level. See Appendix Table C.14.

[25]To select placebo technologies, I select technologies with adoption rate comparable to the adoption of data technologies.

JavaScript (e.g., to show interactive plugins), and Advertising technologies. In Figure 4, I display the 90% confidence interval on the coefficient associated with the effect of each placebo technology. All placebo technologies yield statistically insignificant results. This evidence corroborates the idea that managers use data technology to extract information from web traffic activity.

Figure 4 about here

Decreasing Returns to Data. A natural prediction of managers learning investors' preferences from web traffic data is that giving fund managers additional (new) information should increase the beneficial effect of data. Assuming different data technologies allow to generate different signals, I can test this prediction using the number of data technologies installed by the fund manager. For example, one technology might allow to observe the average time visitors spend on each website's page. In contrast, other technologies might allow to observe information about age, geolocation, and other demographics. Those technologies provide different signals to the fund managers. In columns (1) to (3) of Table 9, I test this prediction.

Table 9 about here

As expected, the effect of data technologies on the ability to attract flows increases with the number of technologies installed. When the squared number of technologies is included in the specification, its coefficient enters negatively (although not significantly) in the regression (column (2)), suggesting concave returns to data.

Competing for Data. Finally, if results are indeed driven by managers extracting information from web traffic activity, the effect of data technologies should decline as more funds adopt such tools. Intuitively, as more agents observe a signal, the marginal benefit of that signal should decline (Grossman and Stiglitz, 1980). Therefore, I test whether there is a competition effect in the data space. To test this prediction, I define a coefficient that

captures "data competition" within a fund category as:

$$\gamma_{c,t} = \frac{\sum_{i=1}^{N_{c,t}} \text{DATA}_{i,c,t}}{N_{c,t}}, \tag{10}$$

where $N_{c,t}$ denotes the total number of funds in category c, in month t; and the sum in the numerator is the number of funds with data technology in place, in category c at time t. A larger $\gamma_{c,t}$ means that a larger fraction of funds within a fund-category collect signals using data technologies. I test whether funds adopting in less competitive markets have stronger benefits from data. Figure 5 shows the coefficient on $\text{DATA}_{i,t}$ (θ in equation 4), by different values of $\gamma_{c,t}$ at the adoption date. For example, the leftmost bar in Figure 5 represents funds adopting a data technology when competition for data is relatively low $\gamma_{c,t} < 25\%$. According to the assertion that the marginal benefit of data should decline as many acquire similar signals, the coefficient is decreasing in competition (from left to right), with last adopters having insignificant effect on their net flows. Furthermore, the effect of competition marginally affects even funds that already installed data technologies. I include an interaction term with $\gamma_{c,t}$ in regression (4), to capture the additional effect of competition within fund-category after adoption. Results are in Table 9, column (4). The coefficient on the interaction term is negative, suggesting again that competition within fund category reduces the marginal benefit of data.

Overall, these findings are consistent with the hypothesis that fund managers observe signals to predict customers' demand. Data technologies allow managers to collect and analyze data from web traffic activity, which helps attract (retail) capital.

5.1 Alternative Channels

The main findings are consistent with the economic framework in Section 2. According to a learning channel, funds observe signals about investor preferences and deviate from their peers (e.g., specialize) to cater more to customers' demand. However, the results of

funds attracting more capital after installing data technologies and charging higher fees are arguably consistent with alternative explanations. Here, I rule out two main alternative stories. First, one possible explanation might be that adoption of data technologies correlates with managers' ability to generate alpha. If funds earn high risk-adjusted performance after installing a data technology, investors should notice it and increase their inflows to those funds. In Appendix Tables C.12 and C.13, I test this alternative explanation with different measures of alpha and value added (Berk and van Binsbergen, 2015), as well as using the recursive demeaning approach in Pástor, Stambaugh and Taylor (2015). Across all specifications, I cannot reject the hypothesis that managers' skills are unchanged after adoption of a data technology. Therefore, since investors should *observe* funds' superior ability to generate performance if this were the main driver of larger inflows, this alternative explanation seems implausible.

Another interpretation of my main findings could be that data technologies are used to *persuade*, rather than learn about, customer preferences (Mullainathan et al., 2008). According to this channel, a mechanism similar to obfuscation and marketing would drive the results on flows and fees. One key implication of this interpretation is that customers would react less to price changes after being subject to persuasion (Varian, 1980; Ellison and Ellison, 2009). To test this hypothesis, I compare the elasticity of fund flows to changes in fees after adopting data technology. If an obfuscation mechanism is at play, the "flows-fee elasticity" should be significantly lower, in absolute terms, after adoption (i.e., flows react less to change in expense ratio while investors are obfuscated). I show that this is not the case in Appendix Table C.15. On the contrary, if anything flows became more elastic to fees change after funds adopt new technology. This result squares with the framework in Section 2: the share of investors who both funds D and N attractive is proportional to the area between $x = 0$ and x_D (shaded blue area in Figure 1). When a fund manager deviates more from $x = 0$, the elasticity of her market share to change in fee will increase. This result is also inconsistent with a "competing for attention" interpretation (Ben-David et al.,

2022), where funds would not compete on fees. Overall, my results are in line with a channel based on fund managers using new technologies to learn about investor preferences. Even though I cannot completely rule out that other mechanisms explain my results, any other interpretation should be consistent with all the results above.

6 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 are concentrated in retail share classes, decline with competition in data collection, and allow funds to maintain a smaller cash buffer when investment opportunities are available, consistent with managers extracting valuable signals about investors' preferences. Moreover, fund managers expand their product offerings after adoption and increase fees, suggesting that customers' data helps optimize both product placement and pricing strategies. To address endogeneity concerns, I exploit the release of TensorFlow, an open-source machine learning library, as positive shock in signals' precision that asset managers can extract from investors' data.

These findings underline the economic importance of asset managers learning customers' preferences to attract capital. This mechanism raises important questions about market efficiency and investors' welfare. While better knowledge of investor preferences could help matching funds to customers, it may also disproportionately benefit larger fund families as they are able to collect more extensive datasets. Moreover, product proliferation might harm investors, as households show well-known systematic biases and make mistakes in their economic decisions.

References

Abdulhafedh, Azad (2021) "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *Journal of City and Development*, Vol. 3, No. 1, pp. 12–30.

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.

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, p. 103745.

Barro, Robert J. and Xavier Sala-I-Martin (1997) "Technological Diffusion, Convergence, and Growth," *Journal of Economic Growth*, Vol. 2, No. 1, pp. 1–26.

Basak, Suleyman and Domenico Cuoco (1998) "An Equilibrium Model with Restricted Stock Market Participation," *The Review of Financial Studies*, Vol. 11, No. 2, pp. 309–341.

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

Ben-David, Itzhak, Francesco Franzoni, Byungwook Kim, and Rabih Moussawi (2022) "Competition for Attention in the ETF Space," *The Review of Financial Studies*, Vol. 36, No. 3, pp. 987–1042, 08.

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, 05.

Birru, Justin, Sinan Gokkaya, Xi Liu, and Stanimir Markov (2024) "Quants and market anomalies," *Journal of Accounting and Economics*, Vol. 78, No. 1, p. 101688.

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 (2024) "Financial Product Incentives to Differentiate: Evidence from Mutual Funds," *Working paper*.

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.

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.

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

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.

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, 09.

Cong, Lin W., Danxia Xie, and Longtian Zhang (2020) "Knowledge Accumulation, Privacy, and Growth in a Data Economy."

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.

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.

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.

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*.

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*.

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, 10.

Farboodi, Maryam, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran (2024) "Valuing Financial Data," *The Review of Financial Studies*, p. hhae034, 07.

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*.

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

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

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.

Hong, Claire Yurong, Xiaomeng Lu, and Jun Pan (2024) "Fintech Platforms and Mutual Fund Distribution," *Management Science*.

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, 05.

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, 09.

ICI, Investment Company Institute (2023) "ICI Fact Book,"Technical report.

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

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.

Lettau, Martin, Sydney C. Ludvigson, and Paulo Manoel (2024) "Characteristics of Mutual Fund Portfolios: Where are the Value Funds?". NBER Working Paper No. w25381.

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, 12.

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.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008) "Coarse Thinking and Persuasion*," *The Quarterly Journal of Economics*, Vol. 123, No. 2, pp. 577–619, 05.

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

Pastor, Lubos, Robert Stambaugh, and Lucian A. Taylor (2024) "Green Tilts." NBER Working Paper No. w31320.

Pástor, Ľuboš, 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.

——— (2021) "Sustainable investing in equilibrium," *Journal of Financial Economics*, Vol. 142, No. 2, pp. 550–571.

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.

Rossi, Alberto G. and Stephen Utkus (2024) "The diversification and welfare effects of robo-advising," *Journal of Financial Economics*, Vol. 157, p. 103869.

Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei (2020) "Marketing Mutual Funds," *The Review of Financial Studies*, Vol. 34, No. 6, pp. 3045–3094, 09.

Savic, Ana, Goran Bjelobaba, Stefana Janicijevic, and Hana Stefanovic (2019) "An Application of PCA Based K-Means Clustering for Customer Segmentation in One Luxury Goods Company," *UBT International Conference*, No. 189.

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

Shy, Oz (1995) *Testing for Weak Instruments in Linear IV Regression*: MIT Press.

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

Stokey, Nancy (2020) "Technology Diffusion." NBER Working Paper No. w27466.

Thakor, Anjan V. (2020) "Fintech and banking: What do we know?" *Journal of Financial Intermediation*, Vol. 41, p. 100833.

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, 10.

Varian, Hal R. (1980) "A Model of Sales," *The American Economic Review*, Vol. 70, No. 4, pp. 651–659.

Veldkamp, Laura (2011) *Testing for Weak Instruments in Linear IV Regression*: Princeton University Press.

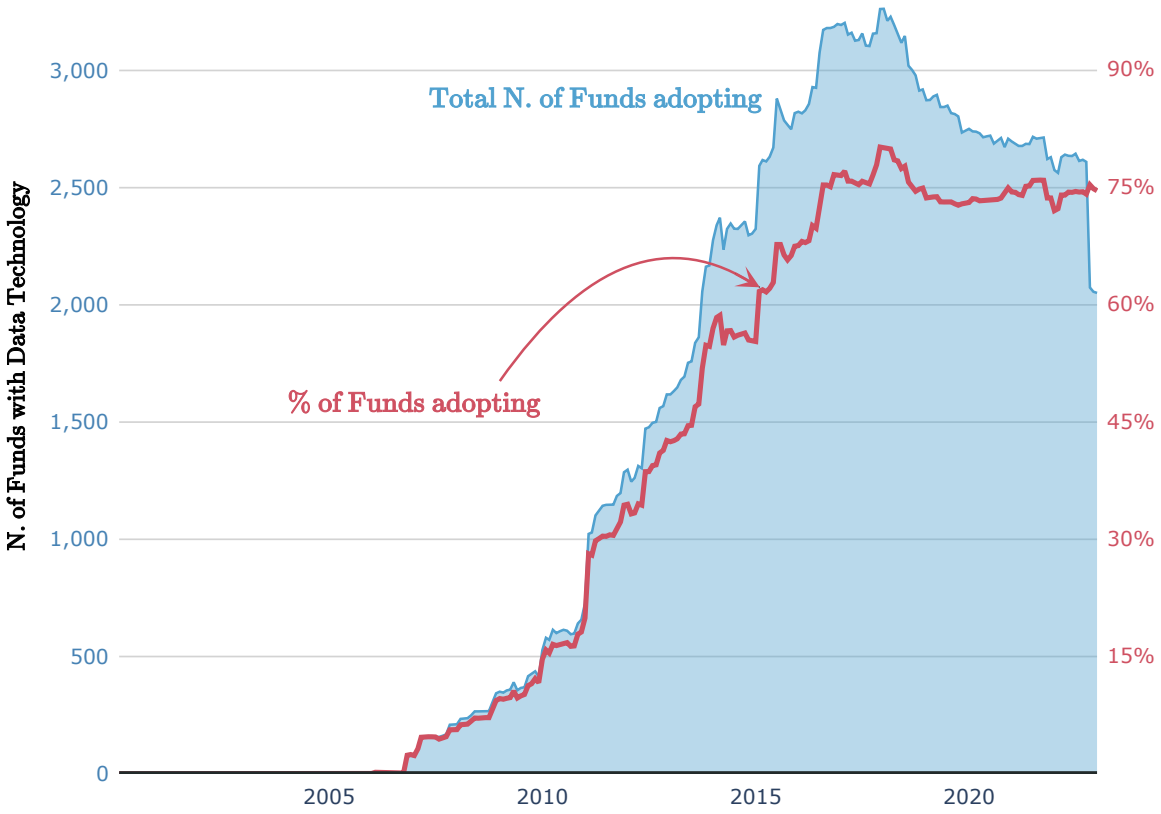


FIGURE 2: **Funds and Data Technology adoption**. This figure shows the adoption of data technologies aimed to capture visitors data, on funds' websites. The data are from BuiltWith, which detects the installation and removal of various technologies by analyzing webpage code. See Section 3.2 for details 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.

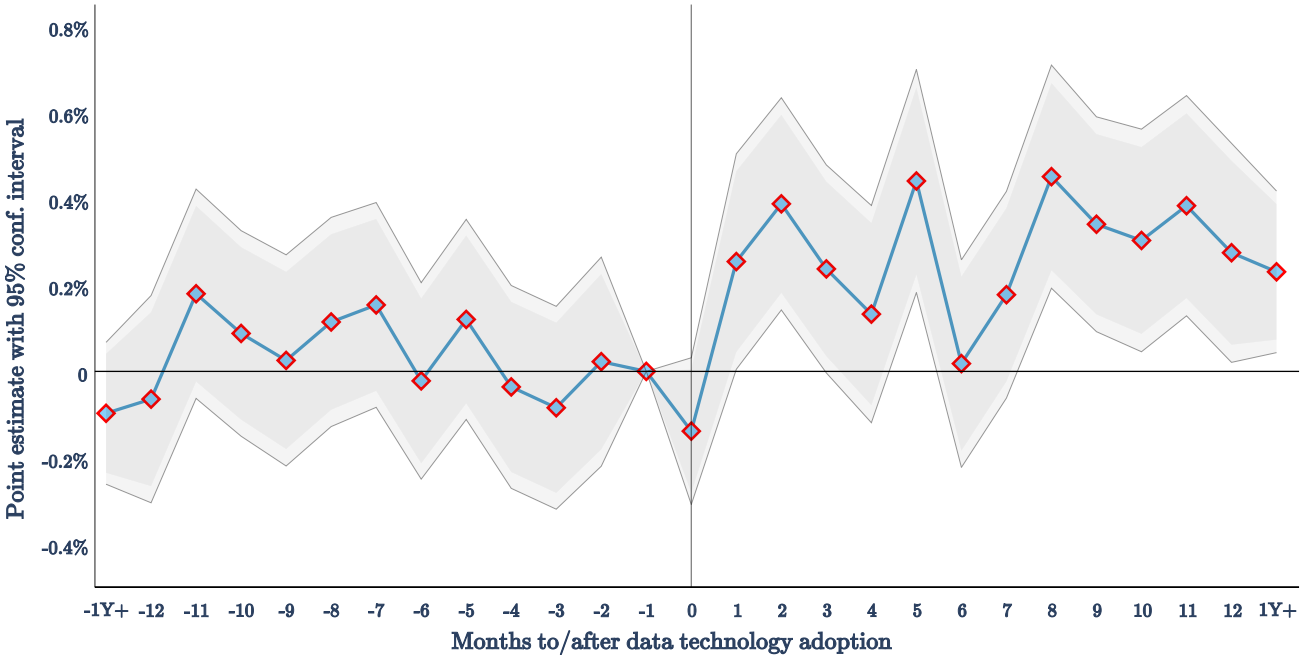


FIGURE 3: **The Dynamic Effect of Data Technologies on Fund Flows**. This figure shows results for the difference-in-differences regression where the dependent variable is the one-month-ahead fund flow. Each point represents the estimated coefficient on the treatment group interaction with each month 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), expense ratio, (log) age, flows, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (10)) 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 rightmost (leftmost) estimates include all observations after (before) 12 months from the adoption month. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022.

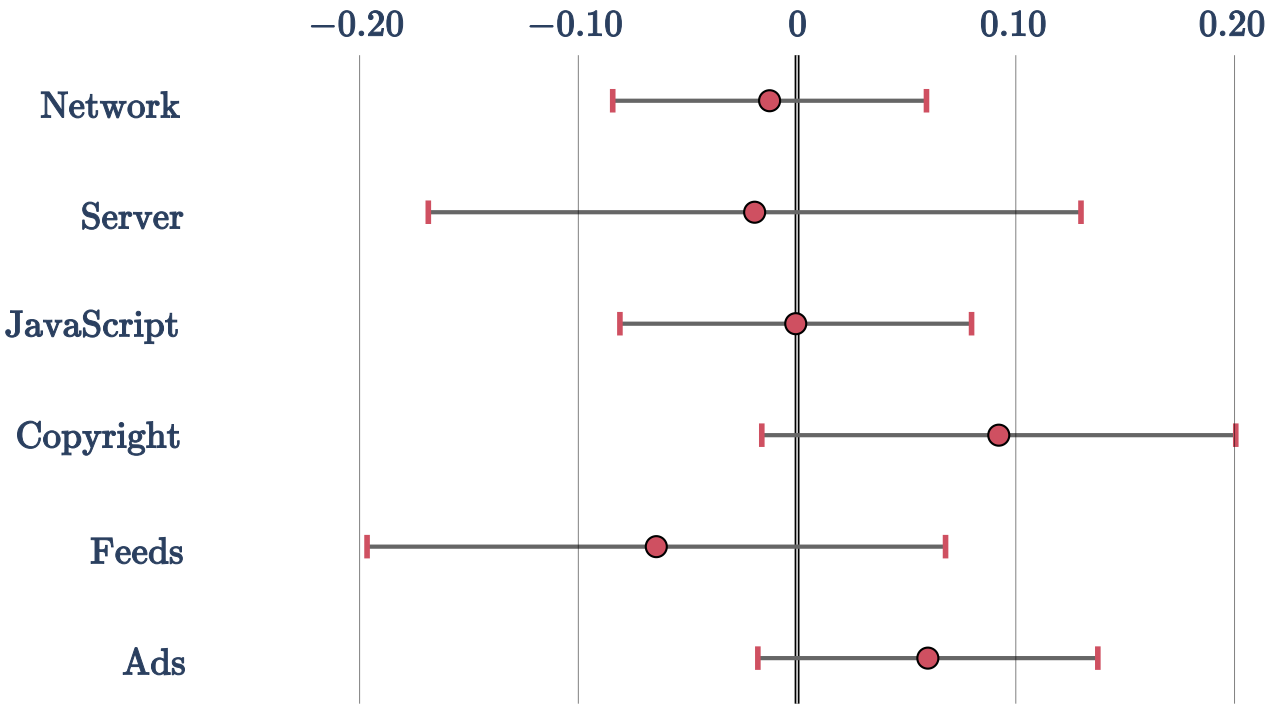


FIGURE 4: **Placebo tests**. This figure shows results from placebo tests on technologies different from data technologies. Each horizontal line represents the 95% confidence interval for tests replacing data technologies with one of the following (placebo) technologies: Network (Content Delivery Network), Server, JavaScript, Copyright, Feeds, and Ads. The specification is the same as the main results in Table 3, column (1). 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 technology of the respective type in place at month t; i.e., analogous to θ in Equation (4). All regressions include fund and time fixed effects, and controls: fund's size (logTNA), (log) age, flows, turnover, and CAPM alpha in month t.

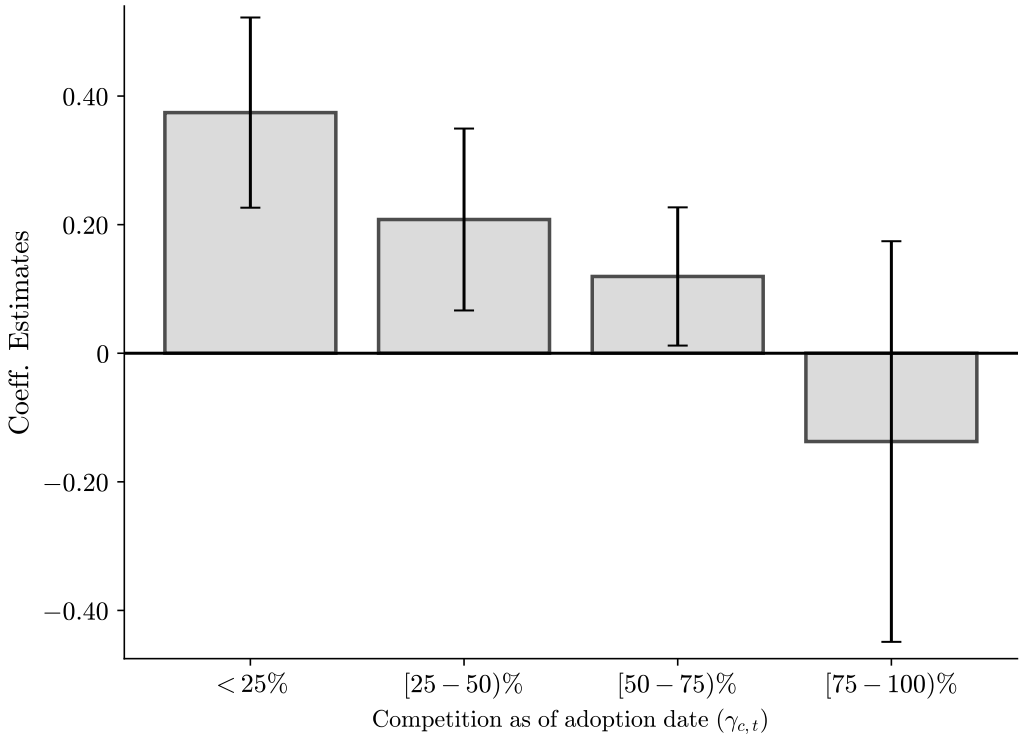


FIGURE 5: **Effect of Competition within Fund-Category**. This figure shows results for difference-in-differences coefficients across different values of competition ($\gamma_{c,t}$) as of data technology adoption. The competition coefficient $\gamma_{c,t}$ is built following equation (10), and it captures 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 (4). All regressions include fund and time fixed effects, and controls: fund's size (logTNA), (log) age, flows, turnover, 12b-1 fees, and CAPM alpha in month t.

	obs.	mean	sd	p5	p25	p50	p75	p95
AUM ($M)	987,242	1,529.46	8,250.46	8.73	54.06	215.77	820.84	5,551.85
Expense Ratio (%)	987,242	1.13	0.53	0.18	0.83	1.12	1.45	2.05
12b-1 Fees (%)	987,242	0.28	0.24	0.00	0.06	0.25	0.41	0.75
Flows (%)	987,242	-0.12	5.71	-6.04	-1.80	-0.63	0.82	7.16
Turnover Ratio	987,242	0.80	1.02	0.06	0.25	0.51	0.95	2.34
Age (Years)	987,242	12.87	8.59	2.50	5.92	11.00	18.25	30.00
Raw Returns	987,242	0.01	0.05	-0.08	-0.02	0.01	0.03	0.08
CAPM Alpha	987,242	0.00	0.10	-0.17	-0.05	-0.00	0.05	0.17
FF5 Alpha	987,242	-0.00	0.26	-0.45	-0.11	0.01	0.11	0.38
N. of *Data Tech.*	987,242	1.09	2.25	0.00	0.00	0.00	1.00	6.00
DATA	987,242	0.35	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 2022.

Data Technology Name	Installation % (in 2022)	Description
Google Analytics	63.50	Users Tracking and Analytics
LinkedIn Insights	37.51	Social Media Tracking and Analytics
Adobe Analytics	29.68	Users Tracking and Analytics
Omniture Test & Target	17.74	A/B Testing
Facebook Pixel	16.98	Social Media Tracking and Analytics
Google Analytics 4	15.04	Users Tracking and Analytics
RapLeaf	13.80	Users Tracking
Twitter Analytics	12.02	Social Media Tracking and Analytics
Bing Universal Event Tracking	11.76	Users Tracking and Analytics
LiveRamp	8.64	Data Connectivity Platform
Yahoo Web Analytics	7.10	Users Tracking and Analytics
Crazy Egg	6.94	Track and Visualize User Interaction
mPulse	6.60	Real Time Customer Experience
Google Optimize 360	6.41	A/B Testing
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 2022. This technologies are aimed to collect and analyze 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 2022. The third column reports a short description of the technology's features.

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.141***	0.127***	0.144***	0.130***
	(0.042)	(0.041)	(0.042)	(0.041)
12b-1 Fees$_{i,t}$			0.251**	0.251**
			(0.112)	(0.109)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.221	0.221	0.221	0.221
Outcome SE	6.293	6.293	6.293	6.293
Obs.	971,730	970,832	971,730	970,832
Adj. R^2	0.112	0.141	0.112	0.141

TABLE 3: **Fund Flows and Data Technologies**: This table shows results of OLS 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$_{i,t}$), 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), expense ratio, (log) age, flows, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. 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$_{i,t}$	0.015***	0.015***	0.018***	0.017***
	(0.004)	(0.004)	(0.004)	(0.004)
12b-1 Fees$_{i,t}$			0.255***	0.237***
			(0.017)	(0.016)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.141	1.141	1.141	1.141
Outcome SE	0.538	0.538	0.538	0.538
Obs.	971,918	971,013	971,918	971,013
Adj. R^2	0.918	0.922	0.921	0.925

TABLE 4: **Expense Ratio and Data Technologies**: This table shows results of OLS panel regression in which 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$_{i,t}$), 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, flows, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$w_{i,t}$(cash) (%)		
	(1)	(2)	(3)
D/P_t	-29.905*		-26.035*
	(15.704)		(15.533)
$\text{DATA}_{i,t} \times D/P_t$		-21.969***	-20.837***
		(4.825)	(4.878)
Controls	✓	✓	✓
Fund FE	✓	✓	✓
Outcome mean	4.367	4.367	4.367
Outcome SE	12.085	12.085	12.085
Obs.	837,070	837,070	837,070
Adj. R^2	0.569	0.569	0.569

TABLE 5: **Portfolio Cash Holdings and Technologies**: This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead portfolio weight in cash. The regressors are the dividend-price ratio (DP_t), an interaction term with 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), expense ratio, flows, turnover, and CAPM alpha in month t. The monthly sample is from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$log(\text{N. of Funds})_{f,t+1}$	N. of Funds $_{f,t+1}$
	(1)	(2)
$\text{DATA}_{f,t}$	0.079**	0.058*
	(0.031)	(0.032)
$log(\text{Age})_{f,t}$	0.3323***	0.479***
	(0.023)	(0.040)
Estimator	OLS	Poisson
Fund Family FE	✓	✓
Time FE	✓	✓
Outcome mean	1.495	13.375
Outcome SE	1.358	30.652
Obs.	159,566	159,566
Adj. R^2	0.906	–
Pseudo R^2	–	0.852

TABLE 6: **Number of Funds in Fund Family and Data Technologies**: This table shows results of OLS panel regression in which the dependent variable is the number of funds offered by fund family f in month $t+1$. 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 ($\text{DATA}_{f,t}$), and the (log) fund family age. See Section 3.2 for details on data technologies. The monthly sample is from March 1993 to December 2022. 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$_{i,t+1}$ (%)						
			z_i :	Tenure of Adoption		N. of Data Tech.	
	(1)	(2)		(3)	(4)	(5)	(6)
DATA$_{i,t}$	0.590***	0.606***	0.654***	0.686***	0.654***	0.647***	
	(0.126)	(0.126)	(0.146)	(0.148)	(0.135)	(0.133)	
DATA$_{i,t}$ × Post_t	0.260**	0.313***					
	(0.109)	(0.109)					
DATA$_{i,t}$ × Post_t × z_i				0.071**	0.091***	0.130***	0.140***
				(0.031)	(0.031)	(0.035)	(0.034)
Controls	✓	✓		✓	✓	✓	✓
Fund FE	✓	✓		✓	✓	✓	✓
Time FE	✓	×		✓	×	✓	×
Category×Time FE	×	✓		×	✓	×	✓
Outcome mean	0.170	0.170		0.170	0.170	0.170	0.170
Outcome SE	6.322	6.322		6.322	6.322	6.322	6.322
Obs.	770,276	769,423	555,493	554,639	662,336	661,458	
Adj. R^2	0.107	0.139	0.096	0.136	0.104	0.137	

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 flow for share class j of fund i. Columns (1) and (2) follow specification in equation (7), while columns (3) to (6) follow (8). 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), expense ratio, (log) age, flows, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (10)) in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022, 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)	(3)	(4)
DATA$_{i,t}$ × Retail	0.467***	0.446***	0.377***	0.358***
	(0.088)	(0.092)	(0.089)	(0.092)
DATA$_{i,t}$	-0.153	-0.165	0.019	0.003
	(0.100)	(0.104)	(0.102)	(0.106)
12b-1 Fees$_{j,i,t}$			0.825***	0.804***
			(0.094)	(0.093)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.691	0.691	0.691	0.691
Outcome SE	10.541	10.541	10.541	10.541
Obs.	799,220	798,719	799,220	798,719
Adj. R^2	0.104	0.114	0.104	0.115

TABLE 8: **Retail and Institutional Share Classes**: This table shows results of OLS 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. All funds in this sample have both retail and institutional share classes, to compare different share classes within the same fund. The control variables include a share class' (*log*) AUM, expense ratio, (*log*) age, flows, turnover, and CAPM alpha in month t. The monthly sample is from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Number of Technologies			Competition Effect	
	Fund Flows$_{i,t+1}$ (%)				Fund Flows$_{i,t+1}$ (%)
	(1)	(2)	(3)		(4)
N. $Tech_{i,t}$	0.030***	0.047***		$\text{DATA}_{i,t}$	0.377***
	(0.007)	(0.018)			(0.114)
N. $Tech^2_{i,t}$		-0.002		$\text{DATA}_{i,t} \times \gamma_{c,t}$	-0.471**
		(0.001)			(0.186)
log(1+ N. $Tech_{i,t}$)			0.130***		
			(0.031)		
Controls	✓	✓	✓		✓
Fund FE	✓	✓	✓		✓
Category×Time FE	✓	✓	✓		✓
Outcome mean	0.221	0.221	0.221		0.221
Outcome SE	6.294	6.294	6.294		6.294
Obs.	850,544	850,544	850,544		971,070
Adj. R^2	0.139	0.139	0.141		0.141

TABLE 9: **Fund Flows, Technology Adoption, and Competition**: This table shows results of OLS 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 column (3). In column (4), 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 an interaction term with the competition coefficient $\gamma_{c,t}$ for fund category c in month t. The competition coefficient is built following equation (10), and it captures the fraction of funds with data technologies in place within fund category-month. 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), expense ratio, (log) age, flows, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Appendix

A Data Appendix

In this Appendix I describe the main dataset's construction procedures. I use four main data sources: (i) CRSP Survivorship-Bias-Free US Mutual Funds data, (ii) Factset Mutual Funds data, (iii) BuiltWith for websites' technology installation/removal dates, and (iv) data from `whois.com` for details on websites' registration date, hosting service, etc.

A.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. CRSP Mutual Funds dataset from January 1980 to December 2022 has 8,803,093 share class-month observations. I start filling missing contact information in CRSP data; i.e., $address1$, $city$, $state$, $website$, and zip. For 62,266 obs. (0.71% of total), the missing information are between two (or more) non-empty information within the same share class, and the two non-empty coincides. I fill those missing observations using the non-empty contact information within the same share class.
CRSP reports a fund's *website* starting January 2008. I use information from `whois.com` to backward fill missing websites' observations. `Whois.com` has information on websites' registration date, hosting service, and others. I hand-collected from `whois.com` the registration date for each website in CRSP (when available), and I verify the website belongs to the CRSP fund using the Internet Archive Wayback Machine. Then, I backward fill 1,423,338 obs. (16.17% of total) missing websites observations before 2008 for months after the website's registration date.

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 836, 636 *cusip* obs. (9.50% 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 fill 1,506,773 obs. (17.35% of total) and 708,187 obs. (8.16% 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*. After this steps, I have 8,237,580 share class-month observations from the CRSP Mutual Funds dataset.

A.2 Factset Mutual 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 (deflated in January 2000 dollars as done for AUM –see Appendix Section A.1).

A.3 CRSP-Factset Merged

I merge CRSP and Factset dataset by *cusip-month*. The CRSP dataset resulting from Appendix Section A.1 has 3,506 observations (0.04% 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.

Then, I merge CRSP and Factset dataset by *cusip-month*. The merge results in 7,646,136/8,237,580 observations merged (92.82%). I drop the remaining 472,797 unmerged observations, since merging them by *ticker-month* might result in incorrect attribution of share classes to fund with re-assigned *ticker*.

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 classify 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 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.

I then aggregate observations across all share classes of the same fund. 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 2,861,642 fund-month observations

A.4 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. 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 A.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/Factset data at the share class level (before aggregation). I merge 3,091,431/7,646,136 observations (i.e., 40.43% of the $crsp_fundno$-*month* observations have at least one data technology in place). Then, I replace missing with zeros, if I have a valid website for the $crsp_fundno$-*month* observation and I have data from BuiltWith for the associated *website-month*, but BuiltWith does not detect data 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 A.1: **Example of website technology data**. This figure displays a snapshot of the website technology data from BuiltWith for `arrowfunds.com` (in my sample).

A.5 Final Sample

From the sample of 2,861,642 $factset_fund_id\text{-}month$ observations, I compute fund flows following Lou (2012):

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

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 dead fund, until five months after. Within this window, I 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 A.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., 1995m5. I flag the 1999m5 observation as merger date. Therefore, following equation A.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 2022.

$crsp_fundno$	$month$	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 A.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 Economic Framework: Proofs

This section contains proofs for the simple framework in Section 2. In $t = 0$, both funds will pick the location on $x \in \mathbb{R}$ that maximizes the number of investors allocating capital to their respective products. In this example, investor preferences are distributed as a logistic distribution, and capturing the larger mass of customers is equivalent to locate in the mode of the distribution. Since $\phi(\mu, \cdot)$ is a symmetric distribution the mode coincides with the mean, and both funds will choos

$$x_j = \mathbb{E}[\mu \mid \mathcal{I}_j], \quad j = \{D, N\}, \tag{B.1}$$

where $\mathcal{I}_j$ is the information set of fund j –i.e., only the prior for fund N, and the prior plus the signal s for fund D.

In $t = 1$, funds compete on fees to maximize their revenues. First, I need to compute the threshold that define the market share of the two funds, i.e., the $\tilde{x}$ such that an investor with preference $\tilde{x}$ will find equivalent buying either of the two funds. This is given by:

$$\tilde{x} \quad \textbf{s.t.:} \quad f_N + t \cdot \tilde{x}^2 = f_D + t \cdot (\tilde{x} - x_D)^2. \tag{B.2}$$

Simple computations yield $\tilde{x} = x_D/2 + \frac{f_D - c}{2x_D t}$, which is equation (1). Now, the optimal fee set by fund D is straightforward:

$$f_D = \text{arg}max\ \pi_D = \text{arg}max\ \underbrace{\Phi(\mu, \tilde{x})}_{\text{Maket share}} \cdot (f_D - c). \tag{B.3}$$

The F.O.C. is:

$$\begin{aligned} 1 - \Phi(\mu, \tilde{x}) &= \phi(\mu, \tilde{x}) \cdot \frac{f_D - c}{2x_D t} \\ 1 - \frac{1}{1 + e^{-(\tilde{x} - \mu)}} &= \frac{e^{-(\tilde{x} - \mu)}}{(1 + e^{-(\tilde{x} - \mu)})^2} \cdot \frac{f_D - c}{2x_D t}, \end{aligned} \tag{B.4}$$

denoting $b := e^{-(\tilde{x}-\mu)}$, and $\kappa := \frac{f_D - c}{2x_D t}$, I can rewrite:

$$\begin{aligned} \frac{1}{1+b} &= \frac{(1+b)^2 - b\kappa}{(1+b)^2} \\ (1+b) &= (1+b)^2 - b\kappa \\ b^2 + b(1-\kappa) &= 0, \end{aligned} \tag{B.5}$$

which gives solution $b_1 = \kappa - 1$, and $b_2 = 0$. As $b := e^{-(\tilde{x}-\mu)}$, $b_2 = 0$ does not yield a real solution. Substituting back b, and κ:

$$\begin{aligned} e^{-(\tilde{x}-\mu)} &= \frac{f_D - c}{2x_D t} - 1 \\ f_D - c &= 2 \cdot \left(e^{-(\tilde{x}-\mu)} + 1\right) \cdot x_D t \\ f_D &= c + 2 \cdot \left(e^{-(\tilde{x}-\mu)} + 1\right), \\ f_D &= c + 2x_D \cdot t \cdot \exp\{-(x_D - \mu)\} \end{aligned} \tag{B.6}$$

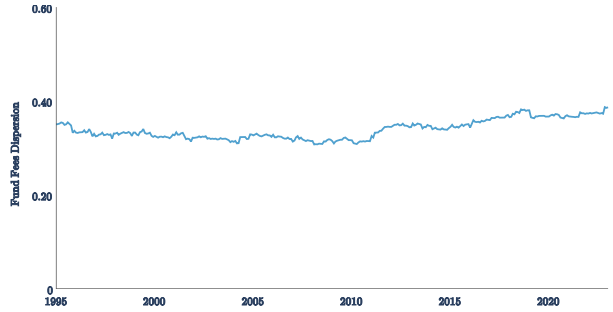
which is equation (2). The S.O.C. for a maximum is satisfied, as:

$$\begin{aligned} &- \phi(\mu, \tilde{x}) \cdot \frac{1}{2x_D t} - \phi'(\mu, \tilde{x}) \cdot \frac{f_D - c}{2x_D t} - \phi(\mu, \tilde{x}) \cdot \frac{1}{2x_D t} = \\ &= -2 \cdot \phi(\mu, \tilde{x}) \cdot \frac{1}{2x_D t} - \phi'(\mu, \tilde{x}) \cdot \frac{f_D - c}{2x_D t} = \\ &= -\phi(\mu, \tilde{x}) \cdot \frac{1}{x_D t} - \phi'(\mu, \tilde{x}) \cdot \frac{f_D - c}{2x_D t} < 0, \quad \text{for } x \in (0, \mu). \end{aligned} \tag{B.7}$$

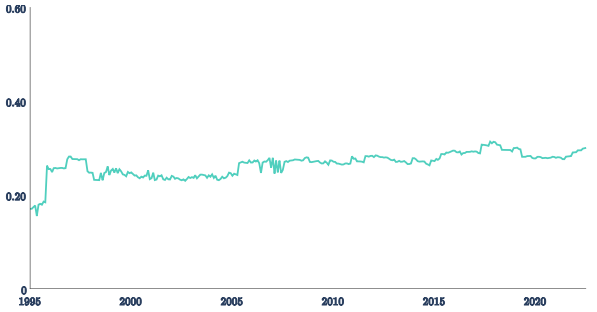
C Additional Results

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

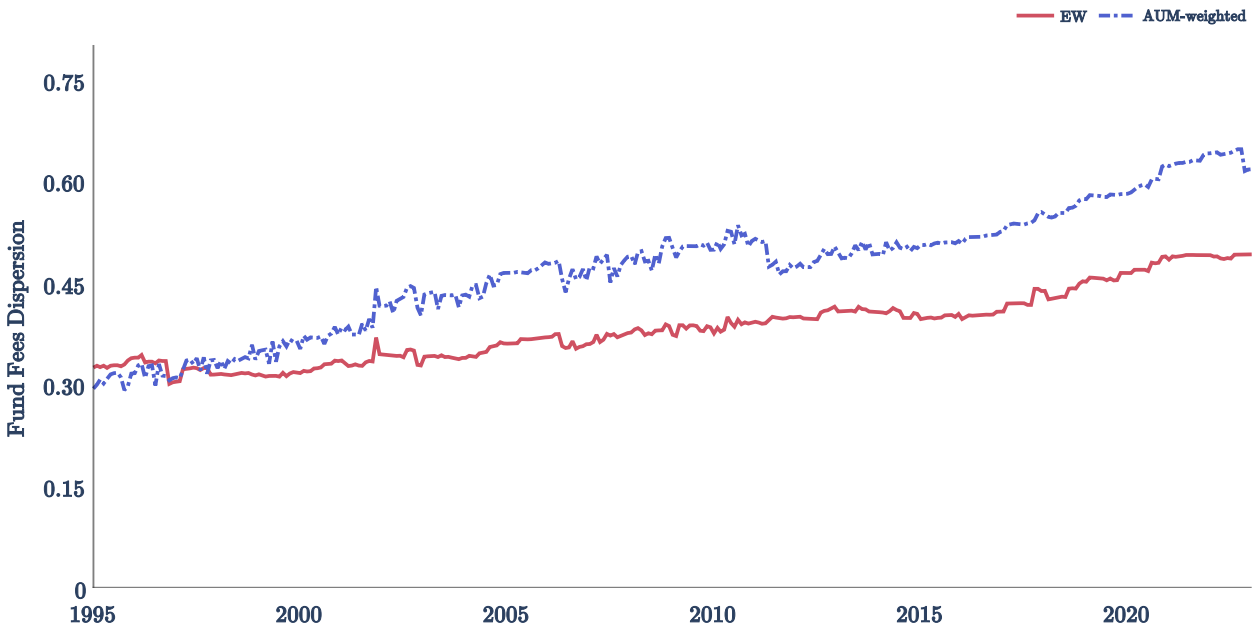
C.1 Appendix Figures



A: Small Cap Funds.



B: Emerging Market Funds.



C: All US Funds.

FIGURE C.1: **Fund Fees Dispersion within Fund Categories**. The figure plots the average fund fees dispersion within fund categories from 1995 to 2022. Panel A and Panel B show the fund fees dispersion (coefficient of variation) for Small Cap and Emerging Market US funds, respectively. Panel C depicts the average fee dispersion within fund category. For each month t, I compute the average and standard deviation of fund fees in a given category. Then, I obtain the average coefficient of variation for each month from 1995 to 2022. The solid red line in Panel C shows the equally weighted dispersion in fund fees, while the dashed purple line reports the AUM-weighted dispersion.

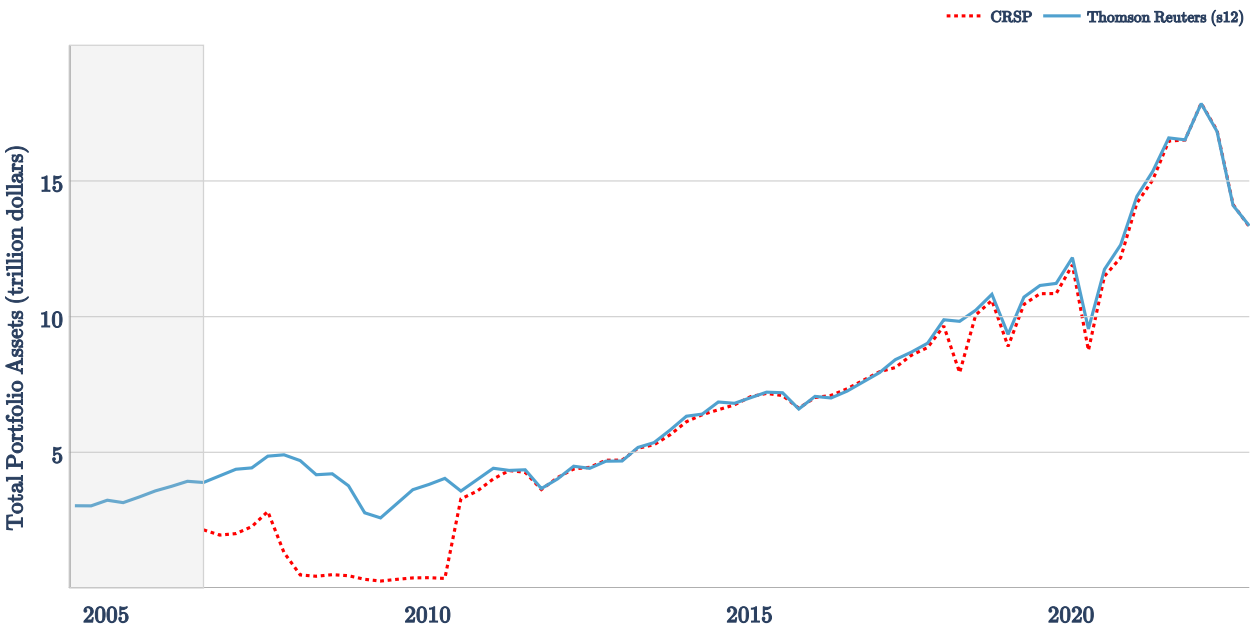


FIGURE C.2: **Total Assets in Thomson Reuters (s12) and CRSP Holdings**. The figure shows the total assets (in trillion of US dollars) in the Thomson Reuters (s12) and CRSP mutual funds holdings databases. CRSP has no holdings data before 2003 (shaded area). The solid blue line reports total equity assets 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.



FIGURE C.3: **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.

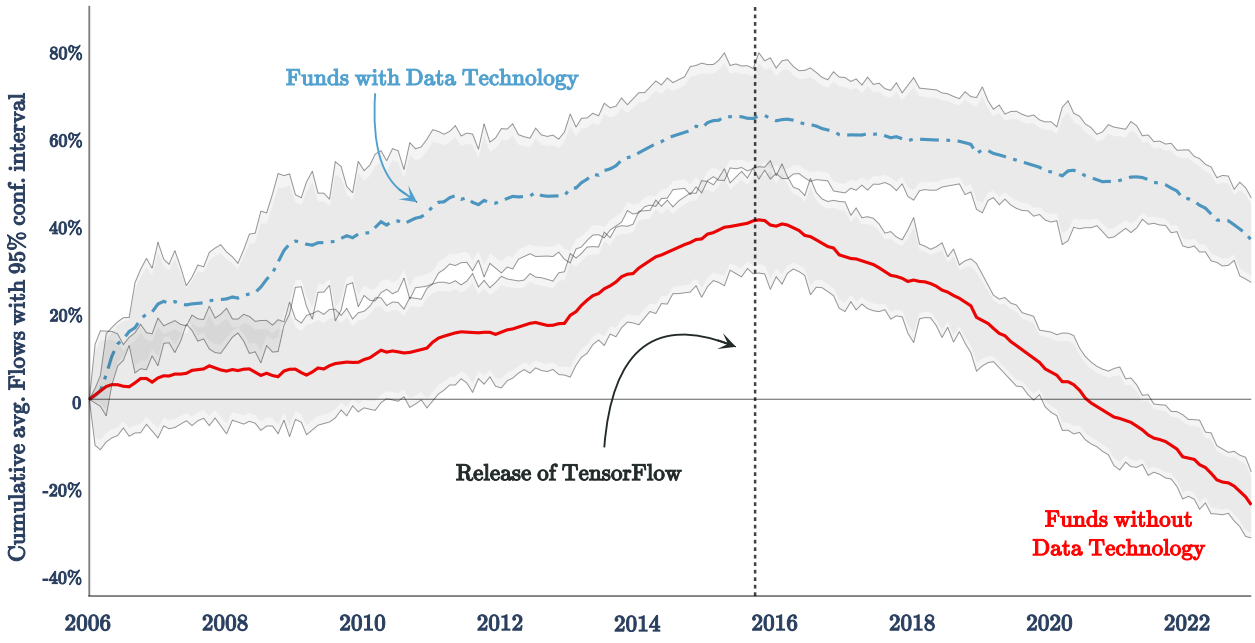


FIGURE C.4: **Cumulative Average Monthly Flows**. The figure shows the cumulative monthly average flows from January 2006 to December 2022, with 95% confidence intervals. For each month t, I compute an equally weighted average of flows for funds with data technology in place (dashed blue line) and funds without data technology (solid red line). See Section 3.2 for details on data technologies. The vertical dotted black line represents the release of TensorFlow in November 2015, which I use as plausibly exogenous variation in signals' precision in Section 4.4. This plot *only* shows sample averages.

C.2 Appendix Tables

	Fund Flows$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.175***	0.173***	0.180***	0.179***
	(0.039)	(0.046)	(0.052)	(0.050)
12b-1 Fees	×	×	✓	✓
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.221	0.221	0.221	0.221
Outcome SE	6.293	6.293	6.293	6.293
Obs.	844,344	830,721	844,345	830,721
Second-stage Adj. R^2	0.001	0.001	0.001	0.001

TABLE C.1: **Fund Flows and Data Technologies, staggered treatment correction**: This table shows results of panel regression robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). I estimate equation (4) following the approach in Gardner (2021); Gardner et al. (2024) to correct for identification concerns in staggered difference-in-differences settings. The dependent variable is the one-month-ahead fund flow, and 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), expense ratio, (log) age, flows, turnover, and CAPM alpha in month t. Columns (3) and (4) include 12b-1 fees as additional controls, following Table 3 in the main text. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are bootstrapped and 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)	(5)	(6)
DATA$_{i,t}$	0.284***	0.220**	0.218***	0.110	0.005	0.007
	(0.053)	(0.087)	(0.083)	(0.115)	(0.097)	(0.113)
12b-1 Fees	×	✓	✓	×	✓	✓
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	×	✓	×	×
Category×Time FE	×	✓	✓	×	✓	✓
Outcome mean	5.698	5.698	5.698	5.459	5.459	5.459
Outcome SE	15.466	15.466	15.466	16.620	16.620	16.620
Obs.	198,791	194,903	194,903	198,790	194,901	194,902
Second-stage Adj. R^2	0.002	0.001	0.001	0.000	-0.000	-0.000

TABLE C.2: **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 (4) substituting the LHS with fund inflows and outflows, following the approach in Gardner (2021); Gardner et al. (2024) to correct for identification concerns in staggered difference-in-differences settings. The dependent variable is the one-month-ahead fund inflows in columns (1) to (3) and fund outflows in columns (4) to (6). 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), expense ratio, (log) age, past flows, turnover, and CAPM alpha in month t. Columns (3) and (6) include 12b-1 fees as additional controls, following Table 3 in the main text. The monthly sample is from January 2006 to June 2018. All standard errors are bootstrapped and 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	-0.190	-0.198	0.008	0.780
Quarterly Fund Flows	-0.241	-0.280	0.039	0.260

TABLE C.3: **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 ($M)	1,509.28	6,934.52	7.27	45.49	197.33	846.95	5,761.43
Expense Ratio (%)	1.06	0.53	0.13	0.74	1.07	1.39	1.97
12b-1 Fees (%)	0.28	0.21	0.00	0.10	0.25	0.39	0.71
Flows (%)	0.30	6.02	-5.54	-1.55	-0.41	1.12	8.17
Turnover Ratio	0.75	1.08	0.05	0.23	0.46	0.85	2.31
Age (Years)	12.48	8.76	1.42	5.08	11.08	18.00	29.67
Panel B: Control							
	mean	sd	p5	p25	p50	p75	p95
AUM ($M)	1,201.05	5,540.89	6.85	40.43	171.23	670.08	4,268.67
Expense Ratio (%)	1.12	0.52	0.18	0.81	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	6.11	-5.76	-1.76	-0.56	1.01	8.76
Turnover Ratio	0.83	1.03	0.07	0.27	0.54	0.99	2.40
Age (Years)	11.58	8.07	1.33	4.83	10.25	16.50	27.42

TABLE C.4: **Balance covariates**: 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.

	Expense Ratio$_{i,t+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.035***	0.033***	0.038***	0.036***
	(0.006)	(0.006)	(0.004)	(0.006)
12b-1 Fees	×	×	✓	✓
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.141	1.141	1.141	1.141
Outcome SE	0.538	0.538	0.538	0.538
Obs.	844,345	830,722	844,344	830,722
Second-stage Adj. R^2	0.012	0.011	0.015	0.013

TABLE C.5: **Expense Ratio and Data Technologies, staggered treatment correction**: This table shows results of panel regression robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). I estimate a similar specification to results in Table 4, following the approach in Gardner (2021); Gardner et al. (2024) to correct for identification concerns in staggered difference-in-differences settings. The dependent variable is the one-month-ahead expense ratio, and 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, flows, turnover, and CAPM alpha in month t. Columns (3) and (4) include 12b-1 fees as additional controls. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are bootstrapped and 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$_{i,t}$	0.149***	0.139***	0.153***	0.149***
	(0.041)	(0.041)	(0.042)	(0.042)
12b-1 Fees$_{i,t}$			0.295**	0.307**
			(0.127)	(0.124)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.349	0.349	0.349	0.349
Outcome SE	6.101	6.101	6.101	6.101
Obs.	683,114	682,065	683,114	682,065
Adj. R^2	0.145	0.176	0.145	0.177

TABLE C.6: **Robustness of Results on Flows, excluding Google Analytics**: This table addresses concerns that the results are entirely driven by Google Analytics. The table replicates the main findings on flows (Table 3) excluding Google Analytics from the set of data technologies. Specifically, this table shows results of OLS 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 (different from Google Analytics) at month t (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), expense ratio, (log) age, flows, turnover, and FF-3 factors alpha (columns (1) and (2)) or FF-5 factors alpha (columns (3) and (4)) in month t. For this table, I remove funds with Google Analytics being the only data technology in place. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. 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$_{i,t}$	0.035***	0.035***	0.046***	0.044***
	(0.007)	(0.007)	(0.007)	(0.007)
12b-1 Fees$_{i,t}$			0.273***	0.252***
			(0.020)	(0.018)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	1.167	1.167	1.167	1.167
Outcome SE	0.560	0.560	0.560	0.560
Obs.	683,114	682,065	683,114	682,065
Adj. R^2	0.928	0.932	0.931	0.935

Table C.7: **Robustness of Results on Expense Ratio, excluding Google Analytics**: This table addresses concerns that the results are entirely driven by Google Analytics. The table replicates the main findings on expense ratio (Table 4) excluding Google Analytics from the set of data technologies. Specifically, this table shows results of OLS panel regression in which 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 (different from Google Analytics) at month t (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, flows, turnover, and CAPM alpha in month t. For this table, I remove funds with Google Analytics being the only data technology in place. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund 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 $\%_{i,t}$	0.171	0.169	-0.283	0.140	0.272	0.314*
	(0.893)	(0.291)	(0.731)	(0.223)	(0.501)	(0.172)
State Adoption $\%_{i,t}$	3.034***	1.145***				
	(0.814)	(0.276)				
City Adoption $\%_{i,t}$			4.066***	1.455***		
			(0.692)	(0.227)		
Zip Code Adoption $\%_{i,t}$					4.766***	1.904***
					(0.463)	(0.183)
Estimator	Logit	Probit	Logit	Probit	Logit	Probit
Controls	✓	✓	✓	✓	✓	✓
Obs.	720,730	720,730	720,730	720,730	720,730	720,730
Pseudo R^2	0.093	0.096	0.141	0.140	0.233	0.226

TABLE C.8: **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), expense ratio, (log) age, and flows. All standard errors are clustered by month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)			
	FF-3 factors		FF-5 factors	
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.149***	0.139***	0.153***	0.149***
	(0.041)	(0.041)	(0.042)	(0.042)
12b-1 Fees$_{i,t}$	0.247**	0.230**	0.232**	0.198*
	(0.113)	(0.110)	(0.114)	(0.112)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.221	0.221	0.221	0.221
Outcome SE	6.293	6.293	6.293	6.293
Obs.	972,199	971,294	972,199	970,832
Adj. R^2	0.106	0.136	0.100	0.129

TABLE C.9: **Robustness of Main Results with alternative performance measures**: This table replicates the main results using different performance measures. This table shows results of OLS 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$_{i,t}$), 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), expense ratio, (log) age, flows, turnover, and FF-3 factors alpha (columns (1) and (2)) or FF-5 factors alpha (columns (3) and (4)) in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Amihud$_{i,q+1}$ (%)			
	(1)	(2)	(3)	(4)
DATA$_{i,q}$	0.496***	0.426***	0.492***	0.433***
	(0.011)	(0.103)	(0.113)	(0.098)
12b-1 Fees	×	×	✓	✓
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.525	0.525	0.525	0.525
Outcome SE	4.588	4.588	4.588	4.588
Obs.	149,870	147,328	149,889	147,346
Second-stage Adj. R^2	0.010	0.007	0.010	0.008

TABLE C.10: **Illiquid Holdings and Data Technologies**: This table shows results of panel regression on the fund portfolio tilt toward illiquid stocks, as measured by Amihud illiquidity. I define portfolio tilts similarly to Pastor et al. (2024). Formally, fund i tilt towards illiquid stocks in quarter q is the absolute value of deviation in weighted characteristics, as defined in Lettau et al. (2024): $\frac{1}{2}\sum_{n=1}^{N}\|(w_{i,q}(n)-\bar{w}_{i,q}(n))\cdot Amihud_q(n)\|$, where $\bar{w}_{i,q}(n)$ is the weight of a value-weighted portfolio within the stocks held by fund i in quarter q. In the estimation, I follow the approach in Gardner (2021); Gardner et al. (2024) to correct for identification concerns in staggered difference-in-differences settings, as the sample of fund holdings data begins in 2004:Q2. The dependent variable is the one-month-ahead fund's portfolio tilt towards illiquid (Amihud) stocks, and 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, flows, turnover, and CAPM alpha in month t. The quarterly sample is from 2004:Q2 to 2022:Q4. All standard errors are bootstrapped and two-way clustered by fund and quarter (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows$_{i,t+1}$ (%)					
z_i :			Tenure of Adoption		N. of Data Tech.	
	(1)	(2)	(3)	(4)	(5)	(6)
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)
DATA$_{i,t}$ × Post_t	0.253**	0.313***				
	(0.109)	(0.109)				
DATA$_{i,t}$ × Post_t × 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 C.11: **Fund Flows and Data Technologies after TensorFlow, without Growth Funds**: This table replicates the findings in Section 4.4 (Table 7 in the main text) removing growth funds from the sample. This robustness address the concern that results in Table 7 are driven by TensorFlow affecting flows to growth funds only. Specifically, this table shows results of OLS panel regression in which the dependent variable is the one-month-ahead flow for share class j of fund i. Columns (1) and (2) follow specification in equation (7), while columns (3) to (6) follow (8). 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 (*log*AUM), expense ratio, (*log*) age, flows, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (10)) in month t. The monthly sample include equity mutual funds and ETFs, excluding growth funds, from March 1993 to December 2022, 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.

	CAPM alpha		FF-3 alpha		FF-5 alpha	
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{DATA}_{i,t}$	0.0003**		0.0002		-0.0000	
	(0.0001)		(0.0001)		(0.0002)	
$\overline{\text{DATA}}_{i,t}$		0.0000		0.0001		-0.0002
		(0.0001)		(0.0002)		(0.0002)
Estimator	OLS	RD	OLS	RD	OLS	RD
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	0.0010	0.0010	-0.0012	-0.0012	0.0012	0.0012
Outcome SE	0.0252	0.0252	0.0363	0.0363	0.0558	0.0558
Obs.	971,918	971,918	971,918	971,918	971,918	971,918
Adj. R^2	0.1199	-0.0067	0.5019	-0.0081	0.6377	-0.0080
First stage F-stat	–	7,421.13	–	7,421.13	–	7,421.13

TABLE C.12: **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), and (5) report results for OLS regressions, while columns (2), (4), and (6) 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, flows, turnover, and performance in month t. I use CAPM alpha (columns (1) and (2)), FF-3 factors alpha (columns (3) and (4)), and FF-5 factors alpha (columns (5) and (6)) as proxy of funds' performance. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Value Added$_{i,t+1}$ (%)					
	CAPM alpha		FF-3 alpha		FF-5 alpha	
	(1)	(2)	(3)	(4)	(5)	(6)
DATA$_{i,t}$	-0.1987		0.0909		-1.5280	
	(0.3315)		(0.8526)		(1.2944)	
$\overline{\text{DATA}}_{i,t}$		-0.0380		0.4459		-0.8166
		(0.3342)		(0.8913)		(1.3246)
Estimator	OLS	RD	OLS	RD	OLS	RD
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	1.294	1.294	-0.887	-0.887	-0.887	-0.887
Outcome SE	34.652	34.652	53.492	53.492	53.492	53.492
Obs.	971,918	971,918	971,918	971,918	971,918	971,918
Adj. R^2	0.0418	-0.0081	0.1421	-0.0074	0.1689	-0.0074
First stage F-stat	–	7,421.14	–	7,421.14	–	7,421.14

TABLE C.13: **Value Added and Data Technologies**: This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead fund value added (see Berk and van Binsbergen, 2015). Value added is defined as fund's (gross) alpha multiplied by AUM. I use CAPM alpha (columns (1) and (2)), FF-3 factors alpha (columns (3) and (4)), and FF-5 factors alpha (columns (5) and (6)) as proxy of funds' performance. Columns (1), (3), and (5) report results for OLS regressions, while columns (2), (4), and (6) 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 (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, flows, turnover, and value added in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. 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$_{i,t+1}$ (%)			
	Retail Funds		Institutional Funds	
	(1)	(2)	(3)	(4)
DATA$_{i,t}$	0.161**	0.190**	0.027	-0.012
	(0.075)	(0.076)	(0.121)	(0.123)
12b-1 Fees$_{i,t}$	-0.598**	-0.635**	-0.139	-0.194
	(0.246)	(0.250)	(0.434)	(0.439)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.421	0.421	0.746	0.746
Outcome SE	7.930	7.930	8.421	8.421
Obs.	407,509	406,755	188,295	185,928
Adj. R^2	0.136	0.161	0.113	0.123

TABLE C.14: **Retail and Institutional Aggregation and Data Technologies**: This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead fund expense ratio. Columns (1) and (2) report results aggregating observations for retail share classes only, while columns (2) and (4) for institutional share classes only. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (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, flows, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. 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)
Exp. Ratio$_{i,t}$	-1.249***	-1.114***	-1.249***	-1.114***
	(0.100)	(0.086)	(0.089)	(0.096)
Exp. Ratio$_{i,t}$ × Data$_{i,t}$		-0.246***		-0.246***
		(0.062)		(0.079)
Controls	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓
Time FE	✓	×	✓	×
Category×Time FE	×	✓	×	✓
Outcome mean	0.222	0.222	0.222	0.222
Outcome SE	6.295	6.295	6.295	6.295
Obs.	971,984	971,079	971,984	971,079
Adj. R^2	0.112	0.141	0.112	0.141

Table C.15: **Fund Flows Semi-Elasticity and Data Technologies**: This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead fund fund flows. The regressors are a the fund expense ratio (in %), a dummy equal to one if a fund i has a data technology in place at month t (Data$_{i,t}$), an interaction term with the fund expense ratio and the dummy, 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, flows, turnover, and CAPM alpha in month t. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2022. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.