

UNIVERSITÀ DELLA SVIZZERA ITALIANA

DOCTORAL DISSERTATION

Data Economy and Financial Markets: Theory and Evidence

Author:

MARCO ZANOTTI

Supervisor:

Prof. Laurent FRÉSARD

Co-advisor:

Prof. Alberto PLAZZI

A dissertation submitted in fulfillment of the requirements

for the degree of

PhD in Economics

SFI PhD in Finance

in the

Institute of Finance

Faculty of Economics

Dissertation Committee:

Prof. Laurent FRÉSARD, Università della Svizzera italiana and SFI

Prof. Alberto PLAZZI, Università della Svizzera italiana and SFI

Prof. Francesco FRANZONI, Università della Svizzera italiana and SFI

Prof. Thierry FOUCAULT, HEC Paris

August 18, 2026

Abstract

The three chapters of my dissertation study how the rise of data availability in the economy affects financial markets and firms. Across industries, firms and investors increasingly adopt new information technologies, such as data analytics and artificial intelligence, to collect and process large volumes of information about their business opportunities, peers, customers, and products. These technologies change what economic agents know, and therefore how they interact with financial markets. In my dissertation I study how new technologies that reduce the cost of making predictions change: (i) how asset managers compete for capital, (ii) how corporate managers use the information contained in stock prices, and (iii) how much information those prices ultimately reveal when *firms* have more data availability.

Chapter 1, *Asset (and Data) Managers*, studies whether asset management companies use customer data to attract capital. Exploiting information from their websites' codes, I track when fund managers begin collecting and analyzing data on their potential customers using tools like Google Analytics or A/B testing. I show that funds adopting such technologies attract 1.5% higher annual flows and charge higher fees, despite no improvement in performance. These results are concentrated in retail share classes and decline with competition, as more rival funds adopt similar tools. At the fund-family level, adopters expand their product offerings, and new funds focus more on retail-oriented themes. Within existing funds, I find evidence of changes in prospectus content and greater sales efforts rather than product differentiation. Overall, data technologies allow managers to raise more capital and charge higher fees, without passing these monetary gains on to investors. These findings show that technological innovation in asset management extends beyond portfolio allocation decisions, and affects how funds attract and retain capital.

In Chapter 2, *Are New Technologies Replacing the Information Produced by Financial Markets?*, co-authored with Laurent Frésard, we turn to retail-oriented firms and ask how data technologies affect their interaction with the stock market. After adopting technologies to exploit customer data collected from their website, firms' investment becomes less sensitive to their stock price. This effect is consistent with a model in which data technologies improve managers' internal information about future demand, and reduce their reliance on stock prices for that information. This "replacement effect" is robust to alternative explanations and reverses when data usage is limited by privacy restrictions. Our findings suggest that the diffusion of data technologies partially replaces the informational role of financial markets in

guiding firms' investment decisions.

Finally, Chapter 3, *Does a Firm's Data Make its Price More Informative?*, studies the direction of learning opposite to the one I study in Chapter 2. While Chapter 2 shows that more informed managers rely less on the information produced by financial markets, Chapter 3 asks how the market's own information changes when firms have more data. In other words, how does firms' adoption of data technologies affect the information contained in their stock prices? To answer this question, I develop a model in which a manager can disclose public information to the market, while investors search for profitable signals in a growing pool of data. More precise internal data unambiguously raises the firm's optimal disclosure. Disclosure, however, attracts derivative content (e.g., media commentary, analyst reports, or social media posts) that dilutes the share of profitable signals and raises investors' cost of producing information. When this externality of public disclosure is strong, price informativeness falls as the firm gains access to more precise data. With multiple firms, adoption at one firm redirects investors' research toward its peers, indirectly reducing their cost of capital. Overall, I show that when firms, rather than investors, have access to more data, the standard prediction that data always makes prices more informative can reverse. This reversal arises from the equilibrium interaction between firms and financial markets, and neither side studied alone delivers it. Together, Chapters 2 and 3 suggest that a data economy affects both what firms learn from markets and what markets learn from firms.

Acknowledgements

I am incredibly indebted to many people that I have been lucky to encounter during my PhD, and these few words will never make justice to the amount of help and support I received during these years. I am grateful to the members of my dissertation committee Laurent Frésard, Alberto Plazzi, Francesco Franzoni, and Thierry Foucault. I am fortunate to have had these primary advisers whose support, advice, and teaching have been invaluable for making me a better researcher and understanding how fascinating our job is. I am especially indebted to Laurent Frésard for guiding me from the beginning, pushing, inspiring, and teaching me what *impact* really is about. I am also grateful to Andrea Tamoni and Alberto Rossi for continuous support, fantastic feedbacks and guidance.

I am also thankful to all former advisers who set the directions for my PhD, namely Roberto Renò, Francesco Rossi and Alessandro Gnoatto at University of Verona, and Alberto Comparini at University of Bergamo. A special thank to a special place, the HEC Paris finance department, for hosting me during one of the most intellectually stimulating period of my PhD, and to the many friends there.

I am thankful to the amazing colleagues I was lucky to have during these years. I thank Roberto, Silvia, Nicola, and Matteo for all the fun, discussion, and learning that I could have asked for during my first years in the PhD. A huge thank to Silvia, for being a good friend, for the stimulating discussions during my entire PhD, and for telling me what I needed to hear even when it wasn't easy to say. I thank Simone, Martina, and Pietro for the support and the joy they have been able to bring from their beginning. We only shared couple of years of our PhD, but felt like they were there for the entire journey. I also thank Pasquale, Yuhan, Hao, Giacomo, Giusy, Nino, Emanuele, Filipe, Dardan, Francesco, Andrea, Shuang, Matteo, Davide, Daniele, and Tina for being such great colleagues.

I have also received fantastic feedback during my PhD from countless amazing people and scholars from all over the world, including but not limited to Andrea Vedolin, Rachel Nam, Francesco D'Acunto, Ahmed Guecioueur, Rich Evans, Carter Davis, Jon Reuter, Olivier Dessaint, Gordon Phillips, Victoria Vanasco, Anthony Lee Zhang, Julien Cujean, Maxime Bonelli, and Ruggero Jappelli.

Last, but not least, I am deeply thankful to a set of people that will never read this manuscript, and yet they have been crucial for everything I have ever achieved: my family. Thank you Federica, my smart, fun, and beautiful life partner, you are the reason I was able to overcome, always with a smile, the difficult moment this

unique journey entailed. You were a constant source of lightness, and I am excited about what the future will bring us in Stockholm or everywhere else we will be. To my parents, Francesca and Claudio, and my brother, Andrea, thank you for always supporting me when I needed and for steering me when necessary, I am deeply grateful. I thank Morena, Piero, and Andrea for their kindness and for welcoming in their family from the first day we met. Finally, I thank Andrea, Diego, Elia, Federico, Sara, Giorgia, Martina, and Eva, for being such good friends and bringing happiness from everywhere in the world.

Contents

Abstract	ii
Acknowledgements	iv
1 Asset (and Data) Managers	1
1.1 Introduction	1
1.2 Theoretical Framework	8
1.3 Data and Measurement	11
1.3.1 Data	11
1.3.2 Data Technologies	13
1.4 Main Findings	17
1.4.1 Funds Flows and Data Technologies	17
1.4.2 Increase in Expense Ratio	22
1.4.3 Endogeneity Concerns	23
1.4.3.1 Instrumental Variable: Data Analytics Graduates . . .	24
1.4.3.2 Shock to Information Precision: Open Source Machine Learning	27
1.4.4 Summary of Additional Robustness Results	30
1.5 Additional Effects of Learning from Customers' Data	31
1.5.1 A Learning Channel	31
1.5.2 Alternative Channels	38
1.6 How do Funds Attract More Flows?	40
1.6.1 Marketing Better Existing Funds	41
1.6.2 Product Differentiation within Fund Family	49
1.6.3 Discussion on Value Added	52
1.7 Concluding Remarks	54
1.8 Appendix	55
1.8.1 Economic Framework	55
1.8.1.1 Equilibrium and Main Predictions	55
1.8.1.2 Value Added	58
1.8.1.3 Endogenous Choice of Adopting Data Technologies . .	59

1.8.2	Data Appendix	62
1.8.2.1	CRSP Mutual Funds	62
1.8.2.2	FactSet Funds	63
1.8.2.3	Morningstar Direct	63
1.8.2.4	CRSP-Morningstar-FactSet Merged	64
1.8.2.5	Website Technologies	65
1.8.2.6	Final Funds Sample	68
1.8.2.7	N-SAR Filings	69
1.8.2.8	Mutual Fund Prospectus Data	70
1.8.2.9	Mutual Funds Portfolio Holdings	71
1.8.3	List of CIP Codes for Instrumental Variable	72
1.8.4	Mutual Funds Prospectuses Text Classification	73
1.8.5	Additional Results	75
1.8.5.1	Appendix Figures	75
1.8.5.2	Appendix Tables	81
References		104

2	Are New Technologies Replacing the Information Produced by Financial Markets?	113
2.1	Introduction	113
2.2	The Replacement Hypothesis	119
2.2.1	The Investment Model	119
2.2.2	The Econometrician's Problem	121
2.2.3	Variation in τ_χ and Identification	123
2.2.4	Endogenous adoption of Data Technologies	124
2.3	Data and Methodology	124
2.3.1	Sample Construction	125
2.3.2	Non-fundamental Shocks to Stock Prices	127
2.3.3	Empirical Specification	131
2.3.4	Data Technologies and Managers' Information: Validation	136
2.4	The Replacement Effect	137
2.4.1	Baseline Results	137
2.4.2	Placebo Tests: Non-data Technologies and Non-retail Firms	145
2.4.3	Regulatory Limitations of Data Access	145
2.5	Alternative Channels and Implications.	148
2.5.1	Learning about the Noise in Stock Prices	148
2.5.2	Fundamental Uncertainty and the Precision of the Price Signal	152
2.5.3	Financing Channel	154

2.5.4	Market Power Channel	157
2.5.5	Implications for Long-run Trends in Investment	157
2.6	Conclusions	160
2.7	Appendix	161
2.7.1	Data Appendix	161
2.7.2	Dividend Reinvestment	163
2.7.3	Appendix Figures	168
2.7.4	Appendix Tables	170
	References	177
3	Does a Firm's Data Make its Price More Informative?	183
3.1	Introduction	183
3.2	Model	189
3.2.1	The Firm	189
3.2.2	The Stock Market	190
3.2.3	Data-Mining Technology	191
3.2.4	Discussion of the Decline in Profitable Signals with τ_{pub}	193
3.3	Equilibrium	194
3.3.1	Market Equilibrium	194
3.3.2	Optimal Disclosure	198
3.3.3	Price Informativeness	200
3.3.4	Cross-Firm Externalities	203
3.4	Extension and Discussion	205
3.4.1	Microfoundation: Sequential Search of Profitable Data	205
3.4.2	Discussion and Implications	209
3.5	Conclusion	211
3.6	Appendix	213
3.6.1	Model Solution	213
3.6.1.1	Information Structure and Beliefs	213
3.6.1.2	Portfolio Choice	214
3.6.1.3	The Linear Rational-Expectations Equilibrium	214
3.6.1.4	The Value of Private Information and Optimal Research	216
3.6.1.5	Comparative Statics	218
3.6.1.6	The Manager's Problem	219
3.6.1.7	Price Informativeness	221
3.6.1.8	The Two-Firm Economy	221
3.6.2	Extension with Sequential Search: Derivations	224
3.6.2.1	The Sequential Search Cost	224

3.6.2.2	The Effective Signal-to-Noise Ratio	225
3.6.2.3	Discussion of The Externality of Public Information .	226
References		228

1 Asset (and Data) Managers

1.1 Introduction

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

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

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

In this paper, I show that asset managers use customer data to improve their ability to attract and retain capital. I exploit information from their websites’ codes and track the adoption of technologies used to analyze customer data, such as Google Analytics or A/B testing tools. These technologies, which I refer to as *data technologies*, allow to

¹See [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.

gather rich information about customers—such as demographics, browsing patterns, and product interest—that managers can use to tailor communication and product design. I find that funds adopting data technologies attract about 1.5% more inflows per year and charge higher fees, despite showing no improvement in performance. The effects are concentrated in retail share classes and, within funds, arise from targeting better potential investors. At the company level, adopters tend to launch more funds in retail-oriented categories, consistent with product differentiation at the margin. Altogether, these results show that asset managers extract value from analyzing customer data and retain all the monetary rents it generates.

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

My empirical strategy relies on information from asset managers' websites to quantify their willingness to collect and analyze investors' data. A key challenge in answering whether asset managers use customer data to attract flows is that we do not observe which managers adopt such practices. To overcome this limitation, I track when asset managers install technologies designed to collect and analyze customers' data. Every website is made of different building blocks, called *technologies*. For example, e-commerce often install technologies enabling secure payments, such as Apple Pay or PayPal. I identify technologies designed to collect and analyze customers' data, such as Google Analytics, and the exact installation date on asset managers' websites. These *data technologies* allow to store digitized information about

user interactions on website pages. For instance, they enable A/B testing (tools similar to Randomized Control Trials), identify which channels bring more potential customers, track users' search history, demographics, and audience overlap with competitors. By observing the adoption date of data technologies by asset managers, I can proxy for when a given manager begins collecting investors' information. This approach provides a novel measure to proxy for the adoption of data technologies that are otherwise unobservable.

I use a staggered difference-in-differences framework to study how asset managers change after installing technologies that process customers' data. I compare changes in monthly flows (pre- and post-adoption) within the same fund and month, controlling for several commonly used covariates. This approach allows me to reduce concerns that time-invariant fund characteristics or common style shocks are driving the results. First, I find that asset managers receive more flows after adopting a data technology on their website. Funds using technologies that analyze investors' information receive approximately 1.5% more flows each year. The magnitude of this effect is not negligible: it is comparable to the increase in flows associated with a fund delivering an annual alpha of 1.75% (the 84th percentile in my sample). In other words, the effect is similar in size to the flow-performance sensitivity for a fund at the 84th percentile of the performance distribution.² Measuring inflows and outflows separately, the result originates from an increase in inflows with no significant change in outflows, implying that funds attract new capital after adoption.

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

I conduct a series of tests to verify that asset managers extract useful information from customer data. Mutual fund data at the share class level offer a clear laboratory

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

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

for studying the heterogeneity of the effects between retail and institutional investors. I find that the effects are concentrated in retail share classes, with no impact on institutional share classes *within* the same fund. These results are consistent with managers learning from customer data, as the technologies I study in this paper are more informative about retail investors.

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

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

Second, I exploit variation in the information that funds can extract from web traffic data, which is plausibly unrelated to fund flows. Specifically, I use the public release of TensorFlow, a widely used open-source machine learning (ML) library in

November 2015.⁴ TensorFlow release improves the precision of prediction models in settings with large data availability. For example, Uber and Airbnb integrated TensorFlow to develop their ML algorithms for rider-driver matching and pricing models, among other things. Intuitively, if the widespread availability of ML allows to extract more informative signals from data, the impact of data technology adoption should intensify after TensorFlow’s release. Consistent with this conjecture, the impact of data technologies is about 30% larger post-release. To tighten identification and validate this mechanism, I exploit cross-sectional heterogeneity in data availability prior to the release of TensorFlow. I construct two proxies for the size of each fund’s dataset at the release date: (i) the number of months between adoption and November 2015, and (ii) the number of distinct data technologies installed. These measures capture the amount of data available to a fund when TensorFlow is released (i.e., (i) proxy for the time series of customer data collected up to TensorFlow’s release, and (ii) proxy for the cross-sectional size of the dataset). Using a difference-in-differences specification with (i) and (ii) as continuous treatment intensity, I find that funds with larger datasets benefit more from TensorFlow’s release. Importantly, these results do not require that all funds actively use ML to analyze customer data; this setting is similar to an intent-to-treat (ITT) design.

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

⁴For reference on the release, see: [wired.com/\[...\]/google-open-sources-its-artificial-intelligence-engine](https://www.wired.com/story/google-open-sources-its-artificial-intelligence-engine/).

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

I consider several alternative explanations for my findings. One rationale behind the results might be that funds adoption of new technologies correlates with a superior ability to generate performance. Although I control for past performance across all specifications and the results on retail share classes (with no effect on institutional share classes within fund) are inconsistent with this hypothesis, I formally test and reject this conjecture using different measures of risk-adjusted performance. Another plausible explanation is that fund managers are not learning about investor preferences but rather persuading them. For instance, adoption could coincide with a rebranding of the fund. I show that this explanation is unlikely in two ways. First, suppose the results reflected funds' rebranding. In that case, they should also hold for other website plugins that are unrelated to data collection, as it is likely that websites rebranding update several plugins simultaneously. However, I find no results using placebo technologies that are not used to collect web traffic data. Second, after adopting data technologies, fund prospectuses become more readable and transparent, inconsistent with obfuscation or persuasion motives (Mullainathan et al., 2008; Ellison and Ellison, 2009; deHaan et al., 2021).

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

Related Literature. This paper contributes to three main strands of literature.

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

Second, the paper contributes to the literature on the industrial organization of the asset management industry.⁵ Hortaçsu and Syverson (2004) show that investors’ search costs are central to explaining why homogeneous S&P500 index funds charge different fees. When investors face search frictions, they choose from a limited subset of available products rather than the best option overall, generating price dispersion even in homogeneous markets. Roussanov, Ruan and Wei (2020) highlight the importance of mutual funds’ marketing for attracting investors’ capital⁶ (Reuter and Zitzewitz, 2006; Kostovetsky and Manconi, 2018; Chen et al., 2022). More recently, Obizhaeva (2024) finds that ETFs attract more flows when advertising through online search engines. According to this strand of literature, technological improvement will reduce search frictions and increase competition among managers. This paper shows a more nuanced role of technology: when information advantages accrue to managers instead of investors, fees rise rather than falling. A related line of literature studies asset managers’ strategic product market choices (e.g., Massa, 2003; Kostovetsky and Warner, 2020; Cvitanic and Hugonnier, 2022; Betermier et al., 2023; Loseto and Mainardi, 2023; Bonelli et al., 2023). In these works, fund families change their product menus to reduce investors’ switching costs or to differentiate themselves. To the best of my knowledge, this paper is the first to investigate whether asset managers collect investors’ information to better meet their demand. In traditional rational models (e.g., Berk and Green, 2004; Berk and van Binsbergen, 2015), learning is one-sided: investors learn about managers’ skill by observing their performance over time. This paper complements these frameworks by showing that learning also

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

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

operates in the opposite direction —managers learn from investors’ data. In that sense, this work bridges the literature on asset management to studies on customer capital (Gourio and Rudanko, 2014; Belo et al., 2014), defined as the stock of a firm’s customer relationship. Roldan-Blanco and Gilbukh (2021) relate firms’ customer capital to markup dynamics⁷, while He, Mostrom and Sufi (2024) show that publicly listed U.S. firms invest more than 4% of their revenue on sales and marketing to build customer capital. This paper shows that advances in data technologies strengthen asset management companies’ ability to build and monetize from customer capital. The same logic may apply to other industries as well.

Third, this paper contributes to the literature on the role of data in the economy (e.g., Brynjolfsson and McElheran, 2016; Goldfarb and Tucker, 2019; Jones and Tonetti, 2020; Cong et al., 2021). Chung and Veldkamp (2024) review this growing literature in detail. The central insight is that the increasing amount of digitized information is valuable for economic agents, and a no-data equilibrium differs from a data economy (Farboodi and Veldkamp, 2023). While data are not conceptually distinct from information, what differs is the enormous number of data points available and the sources from which agents extract that information. A crucial question in understanding the role of data in the economy is whether it increases utilitarian welfare or redistributes rents (Baley and Veldkamp, 2025). I show that funds generate more value added (Berk and van Binsbergen, 2015) after adopting data technologies, and that this additional value originates from retail investors rather than institutional investors. This distinction is crucial for understanding where the surplus from data arises and who captures it. Under strict neoclassical assumptions, the adoption of data technologies is associated with larger utilitarian welfare, as it improves matching between investors and managers. However, if these assumptions do not hold, for example, because of households’ biases, data may redistribute rents from retail investors to fund managers. Overall, I show that valuable data for asset managers are not only datasets for identifying investment opportunities (Farboodi et al., 2021, 2024; Bonelli and Foucault, 2024), but also customers’ data.

1.2 Theoretical Framework

In this section, I describe the main hypotheses regarding the role of data analytics technology for asset managers’ ability to attract capital. I summarize the main intuition here and develop a simple economic framework in Appendix 1.8.1. The

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

theoretical framework builds on Berk and Green (2004) and Berk and van Binsbergen (2015). In these traditional models of competitive markets for asset management services, investors learn about fund managers' skills by observing their risk-adjusted performance over time. However, the learning is one-sided. Whether managers do or do not have information about investor preferences, tastes, or values is irrelevant for the equilibrium of the industry.

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

As in Berk and van Binsbergen (2015), investors allocate wealth across a continuum of funds based on risk and return. In my setting, investors may also have non-pecuniary preferences. These attributes might include a preference for holding green assets (e.g., warm-glow preferences, Hartzmark and Sussman, 2019; Pástor, Stambaugh and Taylor, 2022), or a specific taste over the fund's communication style and frequency⁸ (e.g., daily push notifications, monthly emails, etc.). The precise mix of such preferences evolves slowly over time. Investors know their own tastes, but fund managers do not observe them directly. Instead, managers infer preferences from observing noisy signals.

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

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

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

⁹In a recent interview, Arpit Sarin, Vice President of Digital at Regions Bank, mentioned "understanding how GenZ and TikTok investors pick their investment" and "how much do they trust financial advisors" as urgent questions for which an asset manager "would need a lot of data to get an

Hypothesis 1. (Flows and Data Technologies)

Asset managers collecting more information about investor preferences receive larger fund flows.

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

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

Hypothesis 2. (Expense Ratio and Data Technologies)

Funds with more information about investor preferences charge higher fees compared to competitors.

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

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

answer" ([aiandbanking.libsyn.com/\[...\]/ai-driven-digital-transformation-in-banking](https://aiandbanking.libsyn.com/[...]/ai-driven-digital-transformation-in-banking)).

1.3 Data and Measurement

The data I use in this paper come from multiple sources. This section describes the datasets and summarizes the main cleaning steps; full details are in Appendix 1.8.2. First, I introduce the main sample of US mutual funds and ETFs. Second, I complement these data with information from N-SAR regulatory filings, mutual fund prospectuses, and portfolio holdings. Finally, I describe website technology data and document a series of facts regarding the adoption of data technologies in asset management.

1.3.1 Data

My main data sources are the CRSP Survivorship Bias-Free Mutual Funds dataset, Morningstar Direct, and FactSet Funds. I merge CRSP with FactSet by CUSIP, which identifies unique financial securities and is not re-assigned over time. Then, I merge with Morningstar using CUSIP and ticker. I follow Berk and van Binsbergen (2015) as closely as possible in cleaning funds' data. I outline the main steps here and provide thorough details in Appendices 1.8.2.1 to 1.8.2.4.

The sample consists of US equity mutual funds and ETFs. I adjust all AUM numbers by inflation (in January 2000 dollars) and remove observations prior to a fund's first offer date to mitigate incubation bias concerns (Evans, 2010). I exclude funds with less than two years in the sample and those whose (inflation-adjusted) AUM never exceeds \$5 million (Kacperczyk, Sialm and Zheng, 2008). Funds data are available at the share class level; that is, different share classes belonging to the same fund are reported separately. Therefore, for each month, I aggregate share classes at the fund level by summing the AUM of all subclasses and weighting all other variables (e.g., fees, returns) by lagged AUM. I identify fund families following Dannhauser and Spilker (2023), and estimate fund alphas using 24-month rolling regressions on monthly returns. The sample is from March 1993 to December 2023. The beginning date is March 1993 because earlier AUM data are limited (Pástor, Stambaugh and Taylor, 2015). The final sample comprises 8,125 funds (7,836 equity mutual funds and 289 ETFs), with 947,079 fund-month observations. My sample is somewhat larger than comparable samples of US mutual funds in the literature. The reason is twofold. First, incorporating FactSet yields a more comprehensive merge with CRSP than using Morningstar alone. Second, I do not remove index funds, institutional share classes, sector funds, or funds that allocate less than 80% of their portfolio to stocks.

Following prior literature (e.g., Lou, 2012), I compute the investment flow to fund

i in month t as

$$Flow_{i,t} = \frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t}) - MRG_{i,t}}{AUM_{i,t-1}} \times 100, \quad (1.1)$$

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

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

TABLE 1.1: **Summary Statistics:** This table reports summary statistics for the full sample. For each variable, the table shows the number of available observations (*obs.*), the mean (*mean*), the standard deviation (*sd*), the 5th (*p5*), 25th (*p25*), 50th (*p50*), 75th (*p75*), and the 95th (*p95*) percentiles. AUM is inflation adjusted in January 2000 \$ million. Expense Ratio, 12b-1 Fees, and Flows are in %; e.g., the average fund flow in the sample is -0.12% monthly. The variable *N. of Data Tech.* represents the total number of data technologies installed on the fund's website in a given month, the variable DATA is a dummy equal to 1 if the fund-month observation has at least one data technology installed. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023.

Table 2.3 presents summary statistics for all fund-month observations in my sample. The distribution of AUM is rightly skewed, as is common in the institutional investors' literature. The average expense ratio is 1.12%, with 0.28 percentage points attributable to marketing and distribution expenses (12b-1 fees). On average, funds in my sample have 0.12% negative flows each month and negative net abnormal returns (after fees), consistent with evidence for the U.S.

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

N-SAR filings. Until June 2018, investment management companies filed semi-

annual N-SAR Forms with the SEC.¹⁰ Each Form N-SAR reported fund-level information, but was submitted by the fund complex (i.e., the SEC registrant). A fund complex may include multiple funds. I link all funds within a fund complex to the main CRSP–Morningstar–FactSet dataset (Appendix 1.8.2.7 provides a detailed description of the steps). N-SAR filings contain information not available in traditional datasets, including redemption fees, marketing expenses, captive retail sales forces, 12b-1 fee rebates, and separate reporting of sales (inflows) and redemptions (outflows).¹¹

Fund Prospectuses. Mutual fund prospectuses are from the SEC’s EDGAR (Forms 485APOS and 485BPOS). I closely follow [Abis \(2022\)](#) and [Mullally and Rossi \(2025\)](#) to retrieve prospectuses from EDGAR. The sample period for the prospectus data spans from 2006 to 2023. This sample begins in 2006 because in that year, the SEC started requiring series IDs (fund identifiers) and class IDs (share class identifiers) in their filings. These identifiers enable a direct link to the main sample. For each filing, I parse the text, strip HTML tags, and extract the Principal Investment Strategy (PIS) section ([Abis, 2022](#); [Abis and Lines, 2024](#)), see Appendix 1.8.2.8 for details on the procedure.

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

1.3.2 Data Technologies

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

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

Asset managers are increasingly using these plugins to learn about investor behavior and adapt their products. Vanguard, for example, reports using Adobe Analytics to identify which content is more likely to drive engagement for specific visitors, allowing to target investors more effectively. Similarly, Reliance Mutual Fund

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

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

tracks web traffic to analyze investors' behavior and test alternative website layouts. This practice is typically called A/B testing, and resembles randomized control trials (RCTs) on website pages to run personalization experiments.¹² These plugins provide a natural setting for studying questions about the role of customer data in the asset management industry.

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

I merge BuiltWith data with fund websites in the CRSP-Morningstar-FactSet sample. Fund website information is primarily from CRSP. However, CRSP starts consistently reporting each fund's website only in January 2008. To extend coverage, I hand-collect all website registration dates from `whois.com` and back-fill each fund's website from its registration date to December 2007.¹⁴

Figure 2.1 plots the adoption of data technologies among US funds. The blue area shows the total number of funds with at least one data technology on their website in a given month. The red line reports the share of funds adopting data technologies. A few funds began adopting data technologies as early as 2006. However, the significant surge occurred in 2012, when the adoption rose from 15% to 30%. This surge coincides with a major release of Google Analytics, which remains the leading provider in the space. The adoption of data technologies in asset management continued to grow after 2012 and stabilized around 2019. As of December 2023, approximately 70% of the funds in the sample (over 2,000 unique funds) had adopted at least one data

¹²NIMF Mutual Funds Chief Digital Officer, Arpanarghya Saha, notes that “we need to make our products and services easy to understand —similar to approaches taken at more traditional e-commerce companies.” The firm highlights data technologies as one of its major recent priorities ([business.adobe.com/\[...\]/mutual-fund-case-study](https://business.adobe.com/[...]/mutual-fund-case-study)).

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

¹⁴This procedure marginally increases the number of technologies' adoptions, as most data technologies spread after 2011 (see Figure 2.1). However, this cleaning step ensures I don't misclassify funds as non-adopters before January 2008.

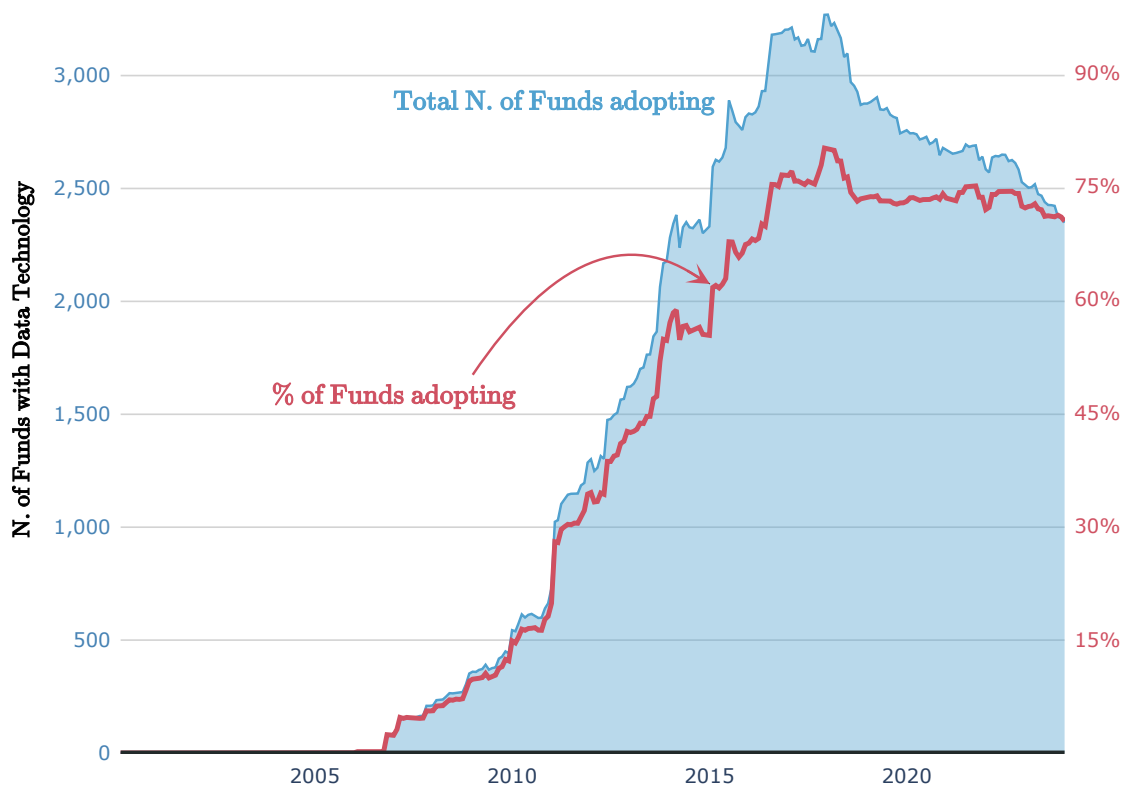


FIGURE 1.1: **Funds and Data Technology adoption.** This figure shows the adoption of data technologies aimed to capture visitors data, on funds' websites. The data are from BuiltWith, which detects the installation and removal of various technologies by analyzing webpage code. See Section 1.3.2 for 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.

technology.

In Table 2.1 I list the leading data technologies, by the end-of-sample. Google Analytics accounts for the lion's share of adoption, with around 60% of funds installing it.¹⁵ Other leading technologies include Omniture Test & Target, which enables A/B testing¹⁶, as well as LiveRamp, which integrates big data across platforms.

All technologies in Table 2.1 are designed to generate signals from web visitors' preferences. These features are not limited to the most commonly used data technologies by asset managers in my sample. Appendix Figure 1.10 shows the word cloud built from descriptions of all data technologies installed by asset managers. Common terms include "tracking", "insight", and "analytics". Overall, these technologies

¹⁵Appendix 1.8.5.2 shows that all the main results are not driven solely by Google Analytics (Appendix Table 1.27).

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

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

TABLE 1.2: **Main Data Technologies:** This table reports the main data technologies installed on funds' websites, as of December 2023. This technologies are allow to collect and process website visitors' data. The second column shows the percentage of funds having the technology installed on its website with respect to the total number of funds, as of December 2023. The third column reports a short description of the technology's features.

suggest that asset managers actively track customers to inform their decisions. The remainder of the paper examines the implications of these technologies for adopters and for the equilibrium of the asset management industry.

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

than one, the websites installed different technologies within hosting-month group.

Appendix Figure 1.9 presents the distribution of cosine similarities within each hosting-month group. The mean similarity is 0.29, with a median of 0.27, both of which are far from one. Websites sharing the same hosting provider install markedly different technologies.¹⁷ Thus, it is unlikely that website technology adoption is determined by the hosting services.

1.4 Main Findings

In this section, I examine how the adoption of data technologies affects asset managers. Guided by the framework in Section 1.2, I first study the impact of data technology on fund flows and fees (expense ratio). I conclude this section with a series of robustness tests in support of the results.

1.4.1 Funds Flows and Data Technologies

The first main hypothesis (Hypothesis 1, Section 1.2) predicts that funds with more data about investor preferences attract more flows. I test this prediction by studying whether funds receive more capital after adopting a data analytics technology on their website. For each fund i and month t , I define a dummy variable $DATA_{i,t}$ equal to one if the fund has at least one data technology in place. Then, I estimate the following specification:

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta DATA_{i,t} + \beta' X_{i,t} + \varepsilon_{i,t+1}, \quad (1.2)$$

where $Flow_{i,t+1}$ is the fund i 's flow between months t and $t + 1$, defined as in equation (1.1). α_i and η_t are fund and time fixed effects, respectively. I also include a specification with fund category×time fixed effects, to mitigate concerns about category-specific shocks. $X_{i,t}$ is a set of control variables, including fund size ($\log AUM$), performance, ($\log$) age, turnover, and 12b-1 fees.¹⁸ I measure performance using the CAPM alpha, as previous research shows it is the closest asset pricing model to what mutual fund investors use (Berk and van Binsbergen, 2016; Barber, Huang and Odean, 2016). Using alternative measures, such as Fama–French 3-factors,

¹⁷For comparison, Panel B of Appendix Figure 1.9 show results from a simulation in which adoption is random within groups. The cosine similarity in the actual data is lower than what would be expected under random assignment.

¹⁸Expense ratio and lagged flows are common controls in fund flows regressions. Yet, both variables are affected by the treatment. As a consequence, including these variables as controls could introduce post-treatment bias. For completeness, I report results with these additional covariates in Appendix Table 1.22. Results are qualitatively unchanged.

5-factors, or alpha with respect to all available Vanguard index funds (Berk and van Binsbergen, 2015) yields the same conclusions (see Appendix 1.8.5).

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

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

TABLE 1.3: **Fund Flows and Data Technologies:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA_{*i,t*}), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund’s size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Table 1.3 shows that data adoption leads to larger fund flows. The first row reports the coefficient of interest, θ . Columns (1)–(2) present baseline difference-in-differences (OLS) results, while columns (3)–(4) show estimates corrected by staggered difference-in-differences concerns (following Gardner et al., 2024). Columns (2) and (4) add fund category×time fixed effects to absorb category-specific shocks. I omit coefficients on control variables for brevity.¹⁹

¹⁹All controls enter with the expected sign across all specifications, e.g., positive flow-performance

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

I report results for both baseline difference-in-differences and estimators addressing heterogeneous treatment effects or other concerns in staggered designs (de Chaisemartin and D’Haultfœuille, 2020; Goodman-Bacon, 2021). The results are similar across all approaches. This is not by chance. In my setting, the total weight attached to “forbidden comparisons” (Goodman-Bacon, 2021) is less than 8%. This relatively low weight arises because the main sample begins several months before the first adoption (treatment), which allows precise estimation of group effects (Gardner, 2021; Gardner et al., 2024). A similar logic applies to period (time) effects. These results remain unaffected by controlling for funds’ Morningstar Rating (Ben-David, Li, Rossi and Song, 2022), using alternative sets of control variables, or removing funds that adopt only Google Analytics. In Section 1.4.4, I thoroughly discuss robustness tests.

In Appendix 1.8.5.2, I examine the parallel trends assumption implicitly behind the identification in equation (1.2), and covariates balance tests. While the parallel trend assumption is, by definition, untestable, I find no evidence of statistically significant differences in pre-adoption trends between treated and controls (Table 1.20). I also cannot reject the hypothesis of covariates balance before adoption at conventional confidence levels (Table 1.21, Figure 1.11).

Figure 1.2 shows the dynamics of the effect on flows. I interact the coefficient of interest with event-time dummies in a stacked sample, following Gormley and Matsa (2011). Then, I report monthly event-time coefficients up to one year before/after adoption, with 95% confidence intervals. The estimates show no significant pre-trends before adoption. The treatment effect is persistently positive only after three months since adoption.

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

sensitivity (Chevalier and Ellison, 1997; Sirri and Tufano, 1998; Pástor et al., 2015; Franzoni and Schmalz, 2017). For comparison with previous works, Appendix Table 1.19 shows detailed results.

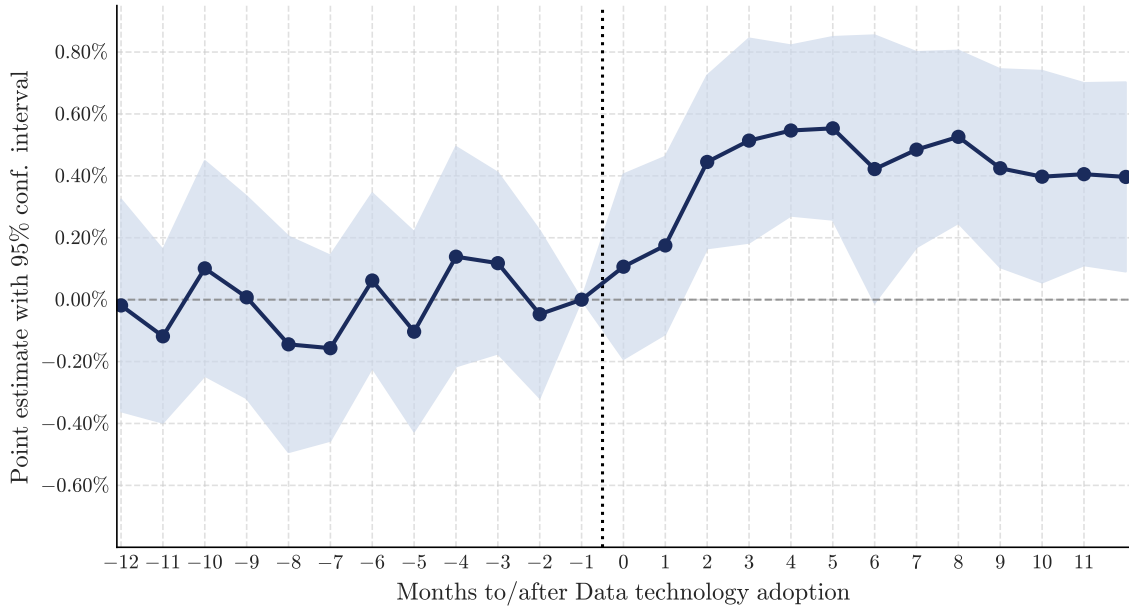


FIGURE 1.2: **The Dynamic Effect of Data Technologies on Fund Flows.** This figure shows results for the stacked difference-in-differences regression where the dependent variable is the one-month-ahead fund flow. Each point represents the estimated coefficient on the treatment group interaction with each month before/after data technology adoption. The treatment is a dummy equal to one if a fund i has a data technology in place at month t ($DATA_{i,t}$). The fund-month control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and alpha with respect to Vanguard index funds (Berk and van Binsbergen, 2015) in month t . Regression include fund and category $\times$ month fixed effects, and the gray area represent the 95% confidence interval for the coefficient estimates. The month just before data technology adoption (-1) is the excluded category in the regression, and is reported as zero in the figure. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023.

The fund flows discussed above (equation (1.1)) represent *net* flows. This measure captures the difference between inflows and outflows for fund i during month t . To better understand the impact of data technology on fund flows, I investigate whether the results stem from larger inflows or smaller outflows. I use data from the SEC's N-SAR filings, which required funds to report their monthly inflows and outflows separately.²⁰ The N-SAR sample spans from January 2006 to June 2018, after which the SEC replaced N-SAR with N-CEN and N-PORT filings. Table 1.4 shows the results for fund inflows and outflows, separately.

Columns (1) and (2) report estimates on inflows, while columns (3) and (4) on outflows. The table shows that the effect of data technology on fund flows arises through larger inflows. The first two columns resonate with previous results: funds that install data technology attract significantly more inflows after adoption. By contrast, the effect on outflows is statistically indistinguishable from zero. Therefore,

²⁰I refer to Appendix 1.8.2.7 for further details on N-SAR filings data.

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

TABLE 1.4: **Fund Inflows, Outflows and Data Technologies:** This table shows results of panel regression on fund inflows and outflows separately, robust to concerns in staggered difference-in-differences (see Goodman-Bacon, 2021). I estimate equation (1.2) substituting the LHS with fund inflows and outflows separately. The dependent variable is the one-month-ahead fund inflows in columns (1) and (2), and fund outflows in columns (3) and (4). All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund’s size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

after adoption, funds receive larger inflows, with no significant difference in outflows. This result suggests that the effect operates mainly on the extensive margin rather than the intensive margin (i.e., funds attracting new investors). This is important because it points to new technologies allowing asset managers to collect more (*new*) capital from investors.

Asset managers ultimately care about scale –that is, the total AUM. While data adoption attracts more flows, there might be some conditions under which this does not translate into larger scale. The increase in flows may be small in dollar terms, short-lived, or offset by weak performance.²¹ To directly verify that adoption increases fund size, I estimate the same specification using the ($\log$) AUM as dependent variable –excluding lagged size from the control variables set to avoid post-treatment bias (Roberts and Whited, 2013). Appendix Table 1.23 report the results. In line with the results on flows, funds that adopt data technologies manage about 12% larger AUM

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

after adoption.

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

1.4.2 Increase in Expense Ratio

The second main hypothesis (Hypothesis 2, Section 1.2) predicts that funds with more customer data charge higher expense ratios in equilibrium. This happens because fund managers align better with investor preferences, beyond the alpha the fund delivers to them. For example, if these technologies allow managers to offer funds closer to investors' taste or values, they can elicit a higher willingness-to-pay and increase fees.

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

TABLE 1.5: **Expense Ratio and Data Technologies:** This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA_{*i,t*}), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

I study this hypothesis with a similar specification to equation (1.2), where the

expense ratio (in %) is on the left-hand side. Table 1.5 reports the results. The estimates imply that funds using data technology increase their fees by approximately 2 to 4 basis points after adoption (first row). Results are similar including category×time fixed effects, and/or other control variables such as the Morningstar rating. Further, the increase in expense ratio is independent of changes in marketing and distribution fees, as I do not find significant changes in 12b-1 fees.

A legitimate concern might be that this increase in fees reflects a mere composition effect within fund. The expense ratio of funds with both retail and institutional share classes is the AUM-weighted average expense ratio across share classes. Because retail share classes are more expensive, a fund's expense ratio could rise mechanically if retail share classes attract more capital than institutional ones. I rule out this explanation, running the specification for a sample of funds with only retail share classes (Appendix Table 1.32).

Theory of the industrial organization of asset management predicts that new technologies will reduce fees by lowering investors' search costs (e.g., Hortaçsu and Syverson, 2004; Gârleanu and Pedersen, 2018; Roussanov et al., 2020 among others). Technological development makes it easier for investors to find and compare asset managers. With easier search, investors contact more managers, identify better funds, and drive the market toward the first-best with no price dispersion and fees equal to marginal cost.²² Thus, these models imply that technological innovations will tend to reduce the cost of financial services. The prediction from Section 1.2 is markedly different and offers a different perspective on the role of new technologies for the equilibrium of the asset management industry. This difference arises because the data technologies I study in this paper facilitate information acquisition for asset managers, not customers. When managers—not investors—benefit from technological innovation, fees rise in equilibrium rather than fall.²³

1.4.3 Endogeneity Concerns

The adoption of data technology may be correlated with unobservable factors. For example, fund managers may install data analytics tools just before an expected surge in product demand. I address this concern using two sources of variation that are plausibly exogenous to fund flows and fees. First, I instrument adoption with the local supply of data analytics graduates. Second, I exploit the release of

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

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

TensorFlow, a widely used machine learning library that improved the precision of signals managers can extract from existing datasets.

1.4.3.1 Instrumental Variable: Data Analytics Graduates

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

The local environment also plays a significant role in asset management. [Christoffersen and Sarkissian \(2009\)](#) show that funds in large cities outperform because managers in financial districts interact more frequently and share investment ideas ([Cujean, 2020](#)). I test whether this intuition extends to technology adoption in asset management. Appendix Table 1.33 shows that funds are more likely to adopt when nearby funds do. To separate geographical effects from peer effects, I regress a fund's adoption decision on the (lagged) share of adopters in its style and in its geographic location. If adoption were driven by within-style peers, the lagged share of adopters in a fund's style would predict adoption. Instead, only the share of adopters in the same zip code, city, or state matters, consistent with a local environment effect. The magnitude of this local effect declines as the geographic unit expands—from zip code to city to state—as in agglomeration economies with technology diffusion ([Rosenthal and Strange, 2004](#)).

When I allow endogenous adoption of a data technology in my economic framework, the prediction echoes intuition from the literature on technology diffusion: Local concentration of experts in data analytics facilitates diffusion by lowering the marginal cost of adoption. Proximity to specialized human capital facilitates learning and access to technology.²⁴ Guided by this idea, I construct an instrument based on the local supply of data analytics graduates. Specifically, I use the number of university graduates in data analytics-related fields from universities near a fund's headquarter. I obtain the annual number of bachelor's, master's, and Ph.D. degrees awarded in data analytics, statistics, and computer science by U.S. universities from the Integrated Postsecondary Education Data System (IPEDS).²⁵ For each year, I aggregate the total number of graduates within a U.S. commuting zone (CBSA) and scale

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

²⁵See Appendix 1.8.3 for the detailed list of Core Instructional Programs (CIPs) in these fields.

by the commuting zone's population. This variable proxies for local concentration of experts that can implement data analysis.

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

The exclusion restriction requires that local data analytics graduates affect fund outcomes only through adoption. This assumption is plausible because fund fixed effects absorb all time-invariant differences across funds, including location choice. A reasonable concern is that local shocks could simultaneously increase graduate supply and investor demand. However, this concern is mitigated by the timing of graduation: completing a degree requires between two to four years, so any local shock that raises enrollment in university programs today would affect graduate supply only with a delay, whereas shocks to investor demand would likely materialize faster. I further address this point scaling the instrument by the CBSA's population, so that it captures talent supply rather than market size. Table 1.6 shows the results.

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

TABLE 1.6: **IV Estimates: Local Supply of Data Analytics Graduates:** This table shows results for the instrumental variable estimates where the instrument is the local supply of graduates in data analytics-related fields. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y , with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix 1.8.3 for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. Columns (5) and (6) show results for the first stage. Columns (1)-(2) and columns (3)-(4) report results for annual fund flows and fees, respectively. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y (DATA $_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in year y . All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

The second stage uses variation from local graduates to estimate the effect of adoption on fund flows (columns (1)-(2)) and fees (columns (3)-(4)). The Kleibergen-Paap statistic to test for weak instruments is around 20 across all specifications, above the threshold of 10 for rejecting weak instruments ([Stock and Yogo, 2005](#)). The second stage results are consistent with earlier evidence. Adoption of a data technology leads to a significant increase in fund flows. At the same time, those funds charge higher expense ratios. On average, plausibly exogenous adoption raises annual flows by 1.7%, (column (2)) and expense ratios by about 3 bps (column (4)). These magnitudes are close to the baseline results in Tables [1.3](#) and [1.5](#). Overall, the IV estimates quantitatively and qualitatively confirm the main predictions, reducing concerns that results purely reflect endogeneity in the adoption.

The results are not driven by big financial districts such as Boston, Chicago, New York, Los Angeles, Philadelphia, or San Francisco. In Appendix Tables [1.34](#) and [1.35](#), I report robustness results removing these financial centers as well as including CBSA×time fixed effects.

1.4.3.2 Shock to Information Precision: Open Source Machine Learning

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

If machine learning raises prediction precision from a given dataset, the effect of

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

data technologies should be stronger after November 2015. To limit concerns that managers adopted a data technology *because* of TensorFlow’s release, I exclude (i) funds that adopt after November 2015 and (ii) funds that adopt within six months before the release.²⁷ This restriction limits endogeneity concerns in adoption: all treated funds in this sample had a technology in place as of November 2015.

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

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 DATA_{i,t} + \theta_2 (DATA_{i,t} \times Post_t) + \beta' X_{i,t} + \varepsilon_{i,t+1}, \quad (1.3)$$

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

Columns (1) and (2) in Table 1.7 report the results. The first row shows estimates for the baseline coefficient on the $DATA$ dummy. Consistent with earlier evidence, unconditionally, funds adopting data technology attract significantly larger flows ($\theta_1 > 0$). The second row reports results for the coefficient of interest in this specification: θ_2 . Consistent with machine learning improving prediction precision, the effect of data technologies is about 30% higher after TensorFlow’s release.

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

$$Flow_{i,t+1} = \alpha_i + \eta_t + \theta_1 DATA_{i,t} + \theta_2 (DATA_{i,t} \times Post_t) + \theta_3 (DATA_{i,t} \times Post_t \times z_i) + \beta' X_{i,t} + \varepsilon_{i,t+1}, \quad (1.4)$$

where z_i is either the tenure of adoption as of November 2015, or the number of technologies installed (i.e., the continuous treatments introduced above). The coefficient of interest is the interaction with the continuous treatment: θ_3 . In both

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

	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*** (0.126)	0.606*** (0.126)	0.619*** (0.143)	0.672*** (0.146)	0.623*** (0.138)	0.643*** (0.139)
$DATA_{i,t} \times Post_t$	0.260** (0.109)	0.313*** (0.109)	0.030 (0.392)	0.001 (0.376)	0.062 (0.143)	0.119 (0.136)
$DATA_{i,t} \times Post_t \times z_i$			0.051* (0.029)	0.077** (0.032)	0.109*** (0.042)	0.113*** (0.039)
Controls	✓	✓	✓	✓	✓	✓
Fund FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Category×Time FE	×	✓	×	✓	×	✓
Outcome mean	0.141	0.141	0.141	0.141	0.141	0.141
Outcome SE	6.308	6.308	6.308	6.308	6.308	6.308
Obs.	770,276	769,423	584,488	583,837	689,466	688,820
Adj. R^2	0.107	0.139	0.094	0.132	0.101	0.133

TABLE 1.7: **Fund Flows and Data Technologies after TensorFlow Release:** This table shows results of OLS panel regression in which the dependent variable is the one-month-ahead fund flow. Columns (1) and (2) follow specification in equation (1.3), while columns (3) to (6) follow (1.4). In columns (3) and (4) the continuous treatment z_i is the ($\log$) number of months between the first data technology adoption and TensorFlow’s release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow’s release, as continuous treatment z_i . $DATA_{i,t}$ is a dummy equal to one if fund i has a data technology in place at month t . See Section 1.3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund’s size ($\log$ AUM), ($\log$) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (1.6)) in month t . The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow’s release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

cases, I expect stronger post-release effects for funds with larger z_i ; equivalently, $\theta_3 > 0$.

Columns (3)-(4) and (5)-(6) in Table 1.7 report results using as continuous treatment the tenure of technology adoption and the number of technologies installed, respectively. The third row shows estimates for the coefficient on continuous treatments. Across all specifications, both continuous treatments confirm the conjecture that the additional effect post-TensorFlow release is stronger for funds with (ex-ante) larger datasets.

Appendix Table 1.37 shows that this specification is robust to excluding growth funds. A concern may be that the release of TensorFlow could raise expectations

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

1.4.4 Summary of Additional Robustness Results

The results above may still reflect variation other than the funds' tracking of customers' data. In Appendix 1.8.5, I run a series of robustness tests to address this concern. I summarize these tests here and report results in the Appendix.

First, Ben-David, Li, Rossi and Song (2022) shows that mutual fund investors rely on simple labels, such as Morningstar ratings, when allocating capital. If data technology adoption correlates with ratings, my results might be spuriously capturing those rating changes. I address this concern controlling for Morningstar rating in Appendix Tables 1.22 and 1.24 (columns (3), (4), (7), and (8)). Results are quantitatively and qualitatively unchanged. This may not appear surprising: Morningstar ratings are assigned based on relative risk-adjusted performance. Thus, mutual funds (and Morningstar) have little discretion over their assignment.

Second, as many fund families host multiple funds on their website (the average is 1.3 websites per family), a reasonable concern might be that the effect comes from a family-specific shock correlated with the adoption (treatment). I address this in three ways. The first approach is similar to how the empirical corporate finance literature deals with treatment at the group-level when multiple firms enter the same group (e.g., multiple firms within a U.S. state in which a law is passed, Giroud and Mueller, 2010; MacKinnon et al., 2023). I cluster standard errors at the *fund family* (group) and time level. In Appendix Tables 1.28 and 1.29, I show that the results are robust to this conservative clustering. In the second approach, I restrict the sample to fund families with only one website per fund (Appendix Table 1.30). Results are unchanged. The third approach is to aggregate observations at the family-month level. Appendix Table 1.31 confirms the results hold.

Third, I use alternative measures of risk-adjusted performance beyond the CAPM alpha. Results are robust and quantitatively stable using different measures of alpha (Appendix Tables 1.25 and 1.26).

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

these controls are standard in the fund flows literature, I test robustness by adding them in Appendix Tables 1.22 and 1.24. Results hold.

Finally, the results are not driven by a unique feature in Google Analytics —the most widely adopted technology in my sample (see Table 2.1). One concern may be that Google prioritizes funds that use Google Analytics in common browser searches, making those funds more visible. I re-estimate the baseline specification excluding Google Analytics from the set of data technologies in Appendix Table 1.27. The results remain robust: estimates are not driven by Google Analytics, but reflect a feature common to all data technologies.

1.5 Additional Effects of Learning from Customers' Data

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

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

1.5.1 A Learning Channel

The findings above hinge on the idea that fund managers learn information from web traffic data. Arguably, web traffic data reveal more information about retail than institutional investors. Thus, the effect of data adoption should be concentrated in retail share classes. Mutual funds data offer a natural setting to test this prediction. I can compare the effect of data technology on retail and institutional share classes *within fund*. I run a specification similar to equation (1.2), but at the share class level:

$$Flow_{j,i,t+1} = \alpha_i + \eta_t + \theta^I DATA_{j,i,t} + \theta^R (DATA_{j,i,t} \times Retail_{j,i,t}) + \beta' X_{j,i,t} + \varepsilon_{j,i,t+1}, \quad (1.5)$$

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

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

TABLE 1.8: **Retail and Institutional Share Classes:** This table shows results of panel regression in which the dependent variable is the one-month-ahead flow for share class j of fund i . The regressors are a dummy equal to one if fund i has a data technology in place at month t ($DATA_{i,t}$) interacted with the share class' j type (retail or institutional), 12b-1 fees, and controls for share class-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. Columns (3) use only retail share classes observations, while columns (4) only institutional share classes observations. The control variables include a share class' ($\log$) AUM, ($\log$) age, flows, turnover, and CAPM alpha in month t . The monthly sample is from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Table 1.8 reports the results and confirms this joint prediction. The effect of adoption is concentrated in retail share classes, without any impact on institutional ones. The coefficient θ^R (second row in columns (1)-(2)), is positive and significant across all specifications. The magnitude aligns with earlier results: data technologies are associated with approximately 2% higher flows per year (0.175% monthly). By

contrast, estimates for institutional share classes are statistically indistinguishable from zero (first row in columns (1)-(2)). Estimates are unchanged when running the specification in equation (1.2) for retail and institutional share classes separately (columns (3)-(4)). These results support the view that website technologies generate signals about retail investors, but provide no information on institutional investors.

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

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

$$\gamma_{c,t} = \frac{\sum_{i=1}^{N_{c,t}} \text{DATA}_{i,c,t}}{N_{c,t}}, \quad (1.6)$$

where $N_{c,t}$ is the total number of funds in category c in month t ; and the sum in the numerator counts funds with data technology in place, in category c at time t . A larger $\gamma_{c,t}$ implies that a larger fraction of funds within a fund-category collect investors’ data. I then estimate specification (1.2) by bins of $\gamma_{c,t}$ at the time of adoption. Figure 1.3 presents the results.

First movers benefit significantly more from adopting data technologies than later adopters. The effect declines monotonically with competition. The leftmost bar in Figure 1.3 shows that funds adopting when competition is low ($\gamma_{c,t} < 25\%$) attract about 5% more annual flows (0.39% monthly). On the other hand, funds adopting when more than 75% of peers already have data show no effect.

Competition also reduces benefits for funds that have already adopted. Appendix Table 1.38 reports regressions with an interaction between adoption and $\gamma_{c,t}$. This term captures how post-adoption competition affects the value of data. The coefficient is negative and significant: benefits from data erode as more competing funds adopt. Estimates suggest that when all peers within a fund’s category analyze customers’ data ($\gamma_{c,t} \rightarrow 100\%$), the additional rents are indistinguishable from zero.

Another natural prediction of a learning channel is that the marginal benefit from adding (new) information should decrease ([Veldkamp, 2011](#)). Different data technologies generate different signals. For example, Hotjar provides heatmaps of web traffic interactions, while LinkedIn Insights allows to collect demographic data

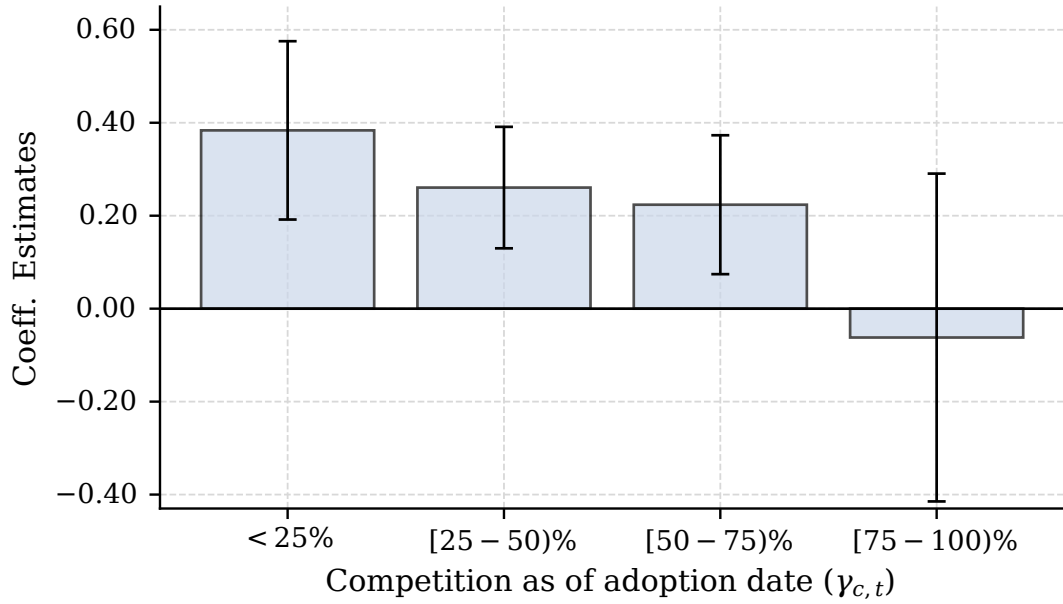


FIGURE 1.3: **Effect of Competition within Fund-Category.** This figure shows results for difference-in-differences coefficients across different values of competition ($\gamma_{c,t}$) as of data technology adoption. The competition coefficient $\gamma_{c,t}$ is built following equation (1.6), and it captures the fraction of funds with data technologies in place within fund category-month. Each bar represents a level of competition as of adoption date (e.g., the first bar represents all fund managers installing their first data technology when less than 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 (1.2). All regressions include fund and time fixed effects, and controls: fund's size ($\log TNA$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t .

from social media. Although these signals differ, their incremental value should fall if managers learn from data.

I can test this prediction using the number of technologies installed by a given fund. I group the number of technologies into five bins: $K = 1$, $K = 2$, $K = [3, 10]$, $K = [11, 15]$, and $K > 15$, where K denotes the number of technology installed. Then, I plot the bin coefficients relative to $K = 0$ (the baseline) in Figure 1.4. Each point estimate shows how flows differ for funds in that bin versus funds with no data technology.

Point estimates rise with K , as a fund adopts more data technologies, but at a decreasing rate. I also overlay a concave line fit, $\hat{y} = \alpha + \beta \log(1 + K)$ (solid blue line). This pattern is overall consistent with decreasing marginal benefits from data. In Appendix Table 1.39, I also estimate a specification including the number of data technologies and its square. When the squared number of technologies is included, its coefficient enters negatively (although not significantly) in the regression, suggesting again concave returns to data.

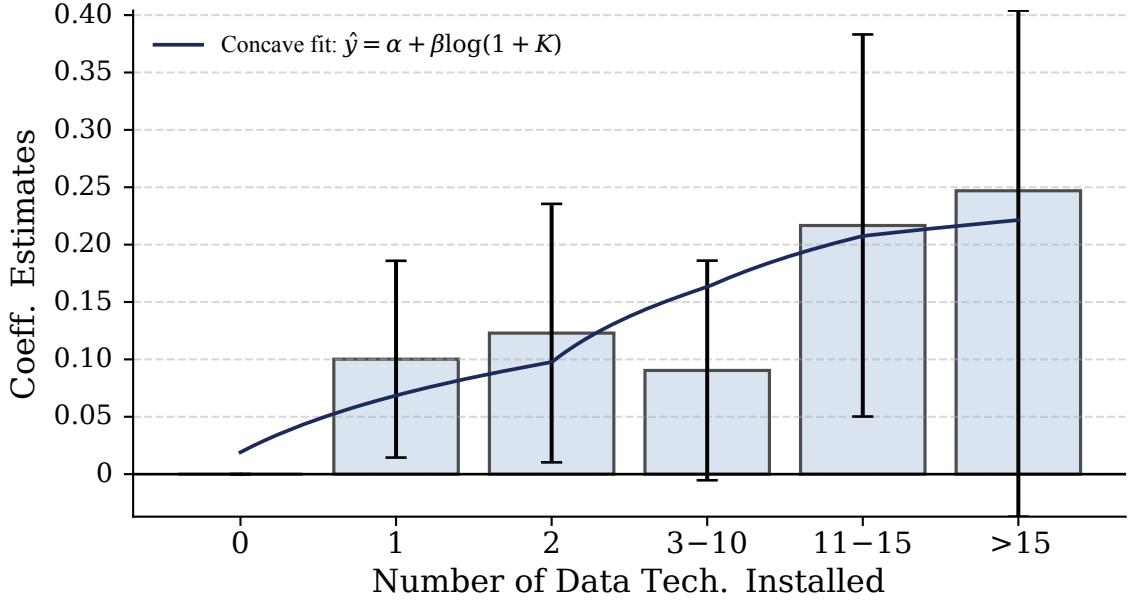


FIGURE 1.4: **Decreasing Marginal Benefits from Data.** This figure shows results for difference-in-differences coefficients across different bins for the number of data technologies installed. Each bar represents a bin in which fund i has K data technologies as of month t , with $K = \{0; 1; 2; [3, 10]; [11, 15]; > 15\}$. Each vertical line represents the 95% confidence interval. The solid line represents a concave fit, estimated with an OLS regression $y = \alpha + \beta \log(1 + K)$. The specification is the same as the main specification in equation (1.2). All regressions include fund and time fixed effects, and controls: fund's size ($\log TNA$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t .

Another implication of learning from data is better liquidity management. Mutual fund managers face uncertainty about investors' redemption needs and therefore hold cash buffers to meet unexpected liquidity demands. Because data reduces uncertainty, managers with more customers' data may better predict redemptions. If so, data managers should hold less cash, particularly when expected returns are high, as holding cash is most costly in such states.

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

$$w_{i,q}(cash) = \alpha_i + \eta_q + \lambda_1 \cdot D/P_{i,q} + \lambda_2 \cdot (D/P_{i,q} \times DATA_{i,q}) + \beta' X_{i,q} + \varepsilon_{i,q}, \quad (1.7)$$

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

Expected returns are high when the dividend-price ratio is high, as equity prices are relatively low compared to dividends. I therefore expect $\lambda_1 < 0$: fund managers, on average, reduce cash to invest in attractive opportunities. The coefficient of

interest in this specification is the interaction term (λ_2), which measures whether data managers reduce cash buffers more aggressively when expected returns are high. As before, identification here comes from time-series variation within funds, not cross sectional differences across funds.

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

TABLE 1.9: **Liquidity Management and Data Technology:** This table shows results of panel regression in which the dependent variable is the quarterly fund's cash holdings (columns (1) to (4)), the portfolio Amihid illiquidity ratio (columns (5)-(6)), and the likelihood that the fund charges redemption fees to investors (columns (7)-(8)). In columns (1) to (4) the regressors are the fund's portfolio dividend-price ratio ($D/P_{i,q}$), an interaction term with a dummy equal to one if a fund i has a data technology in place in quarter q ($\text{DATA}_{f,t}$), and controls for fund-quarter characteristics. See Section 1.3.2 for details on data technologies. The fund-quarter control variables (omitted for brevity) include a fund's size ($\log \text{AUM}$), ($\log$) age, turnover, CAPM alpha, and 12b-1 fees. Portfolio holdings data are from Thomson Reuters (s12). The quarterly sample is from 2004Q2 to 2023Q4. All standard errors are two-way clustered by fund and quarter (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Table 1.9 reports the results. The first two rows present estimates for λ_1 and λ_2 .

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

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

Columns (5) and (6) in Table 1.9 confirm the hypothesis. Fund managers who adopt a data technology increase their exposure to illiquid stocks, as measured by the portfolio’s Amihud illiquidity ratio. Although beyond the scope of this paper, these findings may raise concerns about financial stability, as illiquid assets typically force funds to engage in fire sales when unforeseen shocks occur.

On a similar note, fund managers may be more lenient on redemption fees when they can anticipate redemptions. I can test this last hypothesis directly using N-SAR filings, in which funds disclose semi-annually whether they impose redemption fees to investors liquidating their positions. Columns (7) and (8) in Table 1.9 show that after adoption, funds are about 7% less likely to charge redemption fees to investors ($-0.017/0.240 = -0.07$). This pattern is consistent with data managers being less uncertain about liquidity needs after adoption. Overall, these results support the idea that fund managers learn from customers’ data.

1.5.2 Alternative Channels

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

The first alternative explanation is that the adoption of data technologies correlates with managers’ ability to generate alpha. For example, managers may adopt data

technologies right after acquiring data-analysis expertise. If such skills improve the fund's risk-adjusted performance, investors would notice it and allocate more to that fund. Appendix Tables 1.40 test this hypothesis using several performance measures, as well as using the recursive demeaning approach of Pástor, Stambaugh and Taylor (2015). Across all specifications, I find no material change in fund managers' ability to generate alpha after adoption. Therefore, since investors should be able to observe managers' superior ability if this were the main driver of my results, this alternative explanation seems implausible.

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

To test this alternative, I conduct a series of falsification tests using placebo websites' plugins that are unrelated to customers' data collection. These include, for example, plugins used to improve a website's performance, as well as feeds (used to publish blogs), or advertising technologies. I restrict attention to placebo technologies with adoption rates similar to those of data technologies in my sample, to ensure a fair comparison and that results are not being driven by a lack of power. Then, I re-estimate the baseline regression, replacing $DATA_{i,t}$ with each of the placebo technologies.

Figure 1.5 reports the 95% confidence intervals for all coefficients on placebo technologies. All estimates are statistically indistinguishable from zero. This result contrasts with the positive and significant effects of data technologies. This evidence rules out a rebranding story and reinforces the interpretation that the main findings reflect managers' learning from customers' data.

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

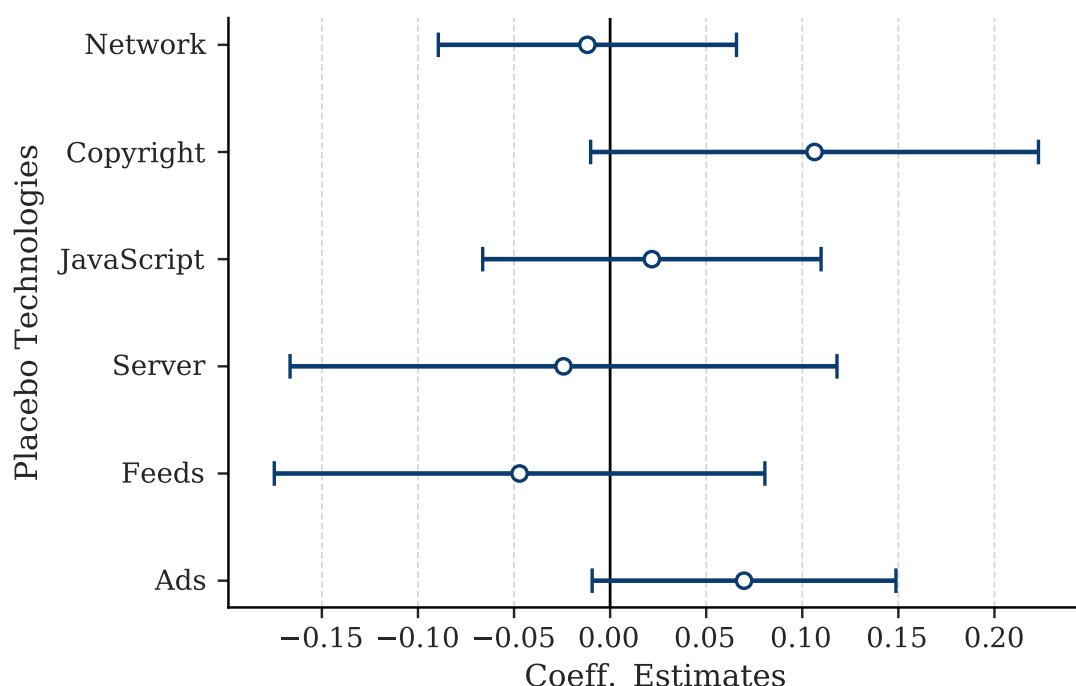


FIGURE 1.5: **Placebo tests.** This figure shows results from placebo tests on technologies different from data technologies. Each horizontal line represents the 95% confidence interval for tests replacing data technologies with one of the following (placebo) technologies: Network (Content Delivery Network), Server, JavaScript, Copyright, Feeds, and Ads. The specification is the same as the main results in Table 1.3. The dependent variable is the one-month-ahead fund flow. The confidence interval refers to the coefficient on a dummy equal to one if fund i has a placebo technology of the respective type in place at month t ; i.e., analogous to θ in Equation (1.2). All regressions include fund and time fixed effects, and controls: fund's size ($\log AUM$), ($\log$) age, turnover, and CAPM alpha in month t .

1.6 How do Funds Attract More Flows?

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

1.6.1 Marketing Better Existing Funds

I first examine the mechanisms at play within existing fund. I compare the effects of data adoption on active and passive funds. Passive funds have little scope for discretionary changes in their portfolio holdings. Therefore, if the effects originate mainly from product differentiation, the results should be concentrated only in active funds. I split the sample into active and passive funds and estimate the main specification separately for each group. I report the results in Table 1.10.

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

TABLE 1.10: **Active and Passive Funds:** This table shows results of panel regression in which the dependent variable is the one-month-ahead flow (columns (1) to (4)) or expense ratio (columns (5) to (8)). In columns (1), (2), (5), and (6) I only include active funds, while columns (3), (4), (7), and (8) only passive funds—including ETFs. All estimates use difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y (DATA $_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund’s size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in year y . The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Pure product differentiation is unlikely to be the only driver of how funds benefit from data technology. I find similarly strong results on flows and fees, for both active and passive funds. The first four columns in Table 1.10 show results on fund flows: columns (1)-(2) for active funds, and columns (3)-(4) for passive funds. In both groups, funds receive significantly more capital after adoption. Columns (5) to (8) confirm the same pattern for the impact of data technologies on fees. All these results are consistent with a selling channel. If product differentiation were the primary mechanism, I should find weak or no effects among passive funds. By

analyzing investors' data, funds improve their targeting and selling strategies rather than solely focusing on product design. This resonates with anecdotal evidence. For instance, [Broadridge \(2023\)](#) describes how fund families can use data to identify which distribution channels are more effective and reallocate resources accordingly.

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

N-SAR filings contain detailed information on funds' marketing activities. In particular, I observe total marketing expenditures as well as the allocation of 12b-1 payments across distribution channels. 12b-1 fees are typically referred to as marketing and distribution fees. They represent annual charges that funds deduct from the AUM to finance distribution activities.²⁸ N-SAR filings report information on how funds distribute their 12b-1 fees payments to different channels, including captive retail sales force, which refers to the *internal* sales force of the fund. Payments to unaffiliated intermediaries must be reported separately. This distinction allows me to distinguish between in-house selling resources from external intermediation.

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

I divide the total dollar amount paid to captive retail sales force by the fund's AUM. Then, I examine whether funds modify their 12b-1 fees payments after adoption. Columns (1)-(2) in Table [1.11](#) show the results. After adoption, managers increase their 12b-1 payments to internal sales resources. This shift toward internal sales is consistent with a selling mechanism that relies more on in-house sales personnel

²⁸12b-1 fees are capped at 1% of AUM annually and are typically reported as part of the fund's expense ratio. Therefore, investors ultimately bear the cost of these fees.

²⁹In Appendix Figure [1.14](#) I examine whether results are concentrated in no-load funds, or they also hold for intermediated (broker-sold) funds ([Del Guercio and Reuter, 2014](#)). Results are similar for both types of funds.

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

TABLE 1.11: **Captive Retail Sales Force:** This table shows results of panel regression on funds' captive retail sales forces, robust to concerns in staggered difference-in-differences (see [Goodman-Bacon, 2021](#)). The dependent variable is the fund's one-month-ahead payment to captive retail sales force from 12b-1 fees, over its AUM (columns (1)-(2)). While in columns (3)-(4) the dependent variable is the fund's total marketing and sales costs over its AUM. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA_{*i,t*}), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

([Kostovetsky and Manconi, 2018](#)).

Appendix Table 1.41 shows how managers reallocate 12b-1 payments across categories. Managers increase spending on internal sales forces by cutting payments to brokers and dealers (external). At the same time, funds redirect resources toward printing and mailing prospectuses to prospective investors (column (4)).

I find similar evidence on broader marketing and sales expenditures. These expenditures differ from the 12b-1 allocations because they cover the total costs of sales, rather than just the portion financed by 12b-1 fees. In practice, funds often spend more than what they collect charging 12b-1 fees, supplementing these with other resources. In columns (3)-(4) of Table 1.11, I use again data from N-SAR filings to examine the change in total marketing expenditures after adoption. Similar to earlier results, total sales expenditures also increase after adoption.

Next, I test whether marketing efforts become more effective after funds adopt data technologies. [Roussanov, Ruan and Wei \(2020\)](#) highlight the importance of mutual funds' marketing for attracting investors' capital (see also [Reuter and Zitzewitz, 2006](#);

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

TABLE 1.12: **Marketing Expenditures and Fund Flows:** This table shows results of panel regression on funds' marketing and sales efforts after adoption of a data analytics technology. The dependent variable is the fund's cumulative six-months-ahead fund flows. The regressors are a fund's (*log*) total marketing and sales expenditures, a dummy equal to one if a fund *i* has a data technology in place at month *t* (DATA_{*i,t*}), the interaction between DATA_{*i,t*} and (*log*) marketing expenditures, and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size (*log*AUM), (*log*) age, turnover, 12b-1 fees, and CAPM alpha in month *t*. The semi-annual sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and calendar semester (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Kostovetsky and Manconi, 2018; Chen et al., 2022 for similar evidence). If data improve the targeting of marketing efforts, I expect a stronger association between marketing expenditures and subsequent fund flows after adoption. Table 1.12 confirms this prediction. The baseline coefficients in columns (1)–(2) (first row) show that marketing and sales spending, on average, positively predicts future flows: a 1% increase in marketing leads to additional 0.19% flows. In columns (3)–(4), I report results for the interaction term of marketing and sales with the adoption of data analytics technology (second row). The relationship between marketing expenditures and flows strengthens after adoption. The interaction term is statistically and economically significant, consistent with marketing and sales efforts becoming more productive after collecting potential customer data.

If funds cater to retail investors, I should find observable changes in their marketing

and distribution material. I test this hypothesis by analyzing whether text in fund prospectuses becomes more appealing to retail investors after adoption. I focus on the Principal Investment Strategy (PIS) section—the key standardized disclosure of a fund’s investment strategy (Abis, 2022). I measure the prospectus readability with the Flesch Reading Ease (FRE) index³⁰ (Flesch, 1948), average words per sentence (deHaan et al., 2021), and by the frequency of second-person pronouns (e.g., “you”, “your”), which prior work in marketing links to stronger retail engagement (Cruz et al., 2017). Table 1.13 presents the results.

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

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

TABLE 1.13: **Text Analysis on Fund Prospectuses:** This table shows results of panel regression in which the dependent variables are from textual analysis of fund prospectuses. Columns (1)-(2) use the Flesch Reading Ease (FRE) index (Flesch, 1948), columns (3)-(4) report results for the average words per sentence in the prospectus deHaan et al. (2021), columns (5)-(6) for the frequency of second-person pronouns (e.g., “you”, “your”), and columns (7)-(8) for the occurrence of words related to themes (Ben-David et al., 2023). I refer to Appendix 1.8.4 for details on the construction of text-analysis measures. The fund-year control variables (omitted for brevity) include a fund’s size ($\log AUM$), ($\log$) age, turnover, CAPM alpha, and 12b-1 fees. Fund prospectuses are from SEC’s EDGAR (see Appendix 1.8.2). The annual sample is from 2006 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

This result is important. The literature on product complexity/obfuscation shows that intermediaries often use complexity to persuade investors and extract rents; here, I find the opposite.³¹ These results help distinguish whether managers use data to meet investors' demand or to persuade them more effectively. The evidence points to a favorable effect of data: Table 1.13 suggests that managers use data to cater to investors' needs (e.g., Gennaioli, Shleifer and Vishny, 2015), rather than to persuade or obfuscate (Célérier and Vallée, 2017).

Fund prospectuses also mention more themes (Ben-David et al., 2023). I measure the frequency at which the prospectus' text includes words linked to themes such as clean energy, cybersecurity, or cannabis (see Appendix 1.8.4 for the entire list of themes and the respective words). Columns (7)-(8) in Table 1.13 show that these changes are again consistent with an effort to tailor toward retail investors' interests.

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

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

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

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

TABLE 1.14: **Product Differentiation for Existing Funds:** This table shows results of panel regression in which the dependent variables are measures of funds product differentiation based on portfolio holdings. Columns (1)-(2) use the holdings distance measure of [Hoberg et al. \(2017\)](#), columns (3)-(4) report results for the portfolio idiosyncratic volatility, and columns (5)-(6) for active share ([Cremers and Petajisto, 2009](#)). The fund-quarter control variables (omitted for brevity) include a fund's size ($\log AUM$), ($\log$) age, turnover, CAPM alpha, and 12b-1 fees. Holdings data are from Thomson Reuters (s12) (see Appendix 1.8.2). The quarterly sample is from 2004Q2 to 2023Q4. All standard errors are two-way clustered by fund and quarter (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

I examine the effects of data technology adoption on product differentiation in Table 1.14. Estimates on $\bar{d}$ are positive, but quantitatively small and marginally significant. These results do not allow me to reject the null hypothesis that funds increase product differentiation after adoption. At best, the evidence points to only limited change in product differentiation.

I find similar evidence using the idiosyncratic volatility of a fund's portfolio returns. The intuition for this measure is simple. If a fund differentiates more, its return should become more idiosyncratic, as it is less spanned by the market or common risk factors. I compute the idiosyncratic volatility of each stock in a fund's portfolio as the standard deviation of residuals from a Fama-French 3-factors model. Then, I aggregate to construct the portfolio idiosyncratic volatility. I find no significant shift in idiosyncratic volatility after adoption (columns (3)-(4) in Table 1.14).

I find again similar evidence on a fund's active share (Cremers and Petajisto, 2009). Active share measures the extent to which a fund deviates from its benchmark. It represents the share of holdings that differ from a fund's benchmark and serves as a standard proxy for how far a portfolio deviates from its benchmark. Columns (5) and (6), in Table 1.14, report no significant change in a fund's active share after adoption. If anything, the effect is mildly negative.³²

Finally, Kostovetsky and Warner (2020) argue that funds perceived by investors as more unique face lower flow-performance sensitivity. Because investors view those products as less substitutable, they are less likely to reallocate capital after observing realization of risk-adjusted performance. Consistent with this idea, in Appendix Table 1.42 I find that after adopting data technologies, funds exhibit a decline in flow-performance sensitivity, suggesting that investors perceive these funds as more unique.

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

1.6.2 Product Differentiation within Fund Family

Fund families with better knowledge of investors' preferences may expand their product offerings to cater to specific customer demands. To test this prediction, I examine whether fund families increase the number of funds they offer after adopting

³²This slight decline is consistent with the prediction of Berk and Green (2004); Berk and van Binsbergen (2015). When flows push a fund's AUM above the scale that maximizes alpha, their framework predicts that managers index the marginal dollar.

a data technology, relative to non-adopters. I aggregate observations at the fund family-month level and estimate the following specification:

$$\log(\text{N. of Funds}_{f,t+1}) = \alpha_f + \eta_t + \delta \text{DATA}_{f,t} + \varepsilon_{f,t+1}, \quad (1.8)$$

where $\text{N. of Funds}_{f,t+1}$ denotes the number of funds offered by family f in month $t + 1$. Similarly to previous specifications, the dummy variable $\text{DATA}_{f,t}$ equals one if at least one fund within family f has adopted a data technology in month t . Family and time fixed effects ensure that identification comes from variation in the number of funds offered before versus after data adoption, relative to the same change for fund families not adopting data technologies.³³ I report results in columns (1) and (2) of Table 1.15.

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

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

TABLE 1.15: **Launch of New Funds:** This table shows results of panel regression in which the dependent variable is the number of funds offered by fund family f in month $t + 1$ in columns (1)-(2). The regressors are a dummy equal to one if at least one fund within family f has a data technology in place at month t ($\text{DATA}_{f,t}$), and the ($\log$) fund family age. See Section 1.3.2 for details on data technologies. In columns (3) to (8), I only use observations on the first prospectus released by *all* newly launched funds in the sample. Columns (3) to (8) regress mentions of words related to retail-oriented themes such as AI, ESG, Cannabis, cybersecurity, political values, space, and video games in the newly launched funds' Principal Investment Strategy (PIS) of their prospectuses. I detail all words I use to identify these themes in Appendix 1.8.4. Columns (3)-(4) consider all thematic words. Columns (5)-(6) consider only AI and Crypto-related words, while columns (7)-(8) uses mentions of ESG words in new funds' prospectuses. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Adopting a data technology leads fund families to increase their product offerings. Columns (1) and (2) of Table 1.15 show that fund families with a data technology offer on average 20% more funds after adoption. In column (2), I control for the fund family's AUM and age, as it is plausible that families develop organizational skills, which reduces the cost of setting up a new fund. The results remain unchanged: fund families expand their product menu after adopting a data analytics technology. This result mirrors the effect of other emerging technologies, such as AI, on firms' product portfolios. Babina, Fedyk, He and Hodson (2024) finds that more AI-intensive firms expand their product varieties, as AI facilitates the accumulation of knowledge and reduces uncertainty in innovation.

I next examine whether the newly launched funds cater to specific investor preferences. I collect Principal Investment Strategy (PIS) text from mutual fund prospectuses, and search for words linked to retail-oriented themes such as ESG, AI, cybersecurity, cannabis, political values, space, and video games.³⁴ Then, I compute the number of such "thematic" words per 100 words of text, and compare newly launched funds by data adopters versus non-adopters. I include year fixed effects to ensure results do not capture the secular rise of thematic funds in recent years (Ben-David, Franzoni, Kim and Moussawi, 2023). Columns (3) and (4) in Table 1.15 confirm this intuition. Funds launched by families with data technology are more likely to reference themes in high demand from retail investors, mentioning 0.07 more theme-related words per 100 words. This effect amounts to a 45% increase from the sample mean of 0.157 theme-related words every 100 words.

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

1.6.3 Discussion on Value Added

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

³⁴Appendix 1.8.4 lists the complete set of keywords.

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

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

Data technologies raise value added through two channels: larger AUM, and higher fees. All else equal, after adoption, each adopting fund attract approximately 1.5% additional flows annually³⁵, compounding to approximately +12% in AUM on average (Table 1.23). This flow effect alone raises value added by about \$600,000 per fund, in January 2000 dollars (about \$1.10 million in September 2025 dollars).³⁶ In addition, funds raise fees by about 3 basis points after adopting data technology. However, this effect is modest, contributing only about \$3,000 per year in value added. Therefore, the increase in value added stems almost entirely from larger AUM.

Crucially, funds do not improve risk-adjusted performance after adopting data technology (see Section 1.5.2). As a result, asset managers capture the entire additional value, rather than sharing it with investors. This surplus represents the monetary value of customers' data accruing to the asset management industry. Across the industry, the additional value added from data technology amounts to approximately \$1.5 billion. Under the neoclassical assumptions discussed above, this surplus coincides with utilitarian welfare gains from a better match between products and investors' preferences.

I emphasize that these conclusions are based on strict assumptions. Therefore, this discussion should be interpreted within this context. For example, this logic

³⁵Fund flows inform about the growth in assets. To evaluate changes in value added, I need to convert these flow effects into their "stock" equivalent (i.e., AUM), reported in Table 1.23.

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

abstracts from whether additional flows come from reallocating funds away from passive strategies or from shifting consumption into savings. Furthermore, if investors' preferences are sub-optimal, data technology may redistribute welfare rather than create new value. In all these scenarios, broader access to customers' data heightens the importance of educating households on investment decisions.

1.7 Concluding Remarks

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

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

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

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

1.8 Appendix

1.8.1 Economic Framework

1.8.1.1 Equilibrium and Main Predictions

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

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

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

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

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

A fraction $\lambda \in (0, 1]$ of asset managers, called “data-managers”, observe an additional independent signal from analyzing customers' data: $s_{D,t} = \theta_t + \eta_{D,t}$, with $\eta_{D,t} \sim \mathcal{N}(0, \tau_D^{-1})$. D stays for data-managers. Data-managers thus combine two independent signals and have total precision $\tau_s + \tau_D$. I treat λ as given here, and endogenize it in Section 1.8.1.3.

Managers update beliefs about θ_t using standard recursive Bayesian updating

(Kalman filter):

$$\begin{aligned}\hat{\theta}_{j,t+1} &= \rho \hat{\theta}_{j,t} + K_{j,t} (s_{j,t} - \hat{\theta}_{j,t}) \\ K_{j,t} &= \rho \frac{\tau_j}{\hat{\Sigma}_{j,t}^{-1} + \tau_j} \\ \hat{\Sigma}_{j,t+1} &= \rho^2 \frac{1}{\hat{\Sigma}_{j,t}^{-1} + \tau_j} + \frac{1}{\tau_\theta},\end{aligned}\tag{1.9}$$

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

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

$$\Omega_{i,t} = \hat{\Sigma}_{j,t}^{-1} = \mathbb{E}[(\theta_t - \hat{\theta}_t)^{-2}].\tag{1.10}$$

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

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

$$\text{TP}_{j,t} = a - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot (\theta_t - \hat{\theta}_{j,t})^{-2}.\tag{1.11}$$

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

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

$$\mathbb{E}[\text{TP}_{j,t}] = 0, \quad \forall j.\tag{1.12}$$

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

$$q_{j,t}^* = \frac{a + \kappa \Omega_{i,t}}{2b}.\tag{1.13}$$

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

$$Flow_{j,t} = \frac{\kappa}{2b}(\Omega_{i,t+1} - \Omega_{i,t}) = \frac{\kappa}{2b}\Delta\Omega_{i,t+1}. \quad (1.14)$$

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

Proof. (Hypothesis 1)

The equilibrium condition (1.12) writes:

$$\begin{aligned} a - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot \mathbb{E}[(\theta_t - \hat{\theta}_{j,t})^{-2}] &= 0 \\ a - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot \Omega_{i,t} &= 0 \\ \rightarrow q_{i,t}(f_{i,t}) &= \frac{a - f_{i,t} + \kappa \cdot \Omega_{i,t}}{b} \end{aligned} \quad (1.15)$$

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

$$\begin{aligned} \max \pi_{j,t} \\ \text{FOC: } 0 &= \frac{a - 2f_{j,t}^* + \kappa \cdot \Omega_{i,t}}{b} \\ f_{j,t}^* &= \frac{a + \kappa \cdot \Omega_{i,t}}{2} \end{aligned} \quad (1.16)$$

This optimum is unique and global. The second order condition $-2/b < 0$ is always satisfied. Finally, substituting (1.16) in (1.15) yields the equilibrium AUM (1.13), from which (1.14) follows directly. $\square$

Hypothesis 2 follows directly from equation (1.15). Profit-maximizing managers charge higher fees, as they better what customers prefer. This result is akin to firms charging higher prices for products that elicit higher willingness to pay. Selling

³⁷Equation (1.14) also implies that stronger decreasing returns to scale (DRS) reduce the flow benefit of data. Appendix Figure 1.12 confirms this prediction. I follow the recursive demeaning approach in Pástor, Stambaugh and Taylor (2015) to estimate fund-specific DRS coefficients as-of adoption of a data technology. Then, I plot the effect of data technologies on flows for different bins of DRS. As predicted by the theoretical framework, the relationship between the effect of data on flows, and DRS is monotonic and decreasing. By contrast, equation (1.16) implies no relationship between data technology and fees. I show this is indeed the case in Appendix Figure 1.13.

specialized products allows firms to command higher prices (Menzio, 2023), a common result in IO several models (Salop, 1979; Lancaster, 1966; Pellegrino, 2024).

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

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

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

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

$$\beta_{Flows, t \rightarrow t+T}^{DiD} = \frac{\kappa}{2b}(\Omega_{A,t+T} - \Omega_{B,t+T}). \quad (1.17)$$

Therefore, the more precise the additional signal A receives, the larger is β_{Flows}^{DiD} . Moreover, equation (1.17) predicts that the diff-in-diff coefficient on flows is smaller for funds facing larger decreasing returns to scale —i.e., funds for which b is large. Appendix Figure 1.12 confirms that this is the case: β_{Flows}^{DiD} decreases monotonically with the decreasing returns to scale parameter b .

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

$$\beta_{Fees, t \rightarrow t+T}^{DiD} = \frac{\kappa}{2}(\Omega_{A,t+T} - \Omega_{B,t+T}). \quad (1.18)$$

Again, a more precise signal from data analytics technologies is associated with a higher β_{Fees}^{DiD} . Unlike flows, however, β_{Fees}^{DiD} does not depend on decreasing returns to scale. Appendix Figure 1.13 confirms that the effect of data adoption on fees is flat across funds facing different decreasing returns to scale.

1.8.1.2 Value Added

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

$$VA_{j,t} = q_{j,t}^* \cdot f_{j,t}^* = \frac{(a + \kappa \cdot \Omega_{i,t})^2}{4b}. \quad (1.19)$$

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

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

$$VA_{j,t}^{FIN} = q_{j,t}^* \cdot \alpha_{jt}^G = \frac{a^2 - \kappa^2 \cdot \Omega_{i,t}^2}{4b}. \quad (1.20)$$

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

$$VA_{j,t}^{DATA} = \frac{a\kappa \cdot \Omega_{i,t} + \kappa^2 \cdot \Omega_{j,t}^2}{2b}. \quad (1.21)$$

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

The *total* value added rises with stock of knowledge, but its composition shifts. The share of value added extracted from financial markets declines with funds' accumulating more data. While the share coming from analyzing data grows. As funds rely more and more on customers' data, data-driven value added crowds out the value managers generates investing in financial markets. Figure 1.6 illustrates this decomposition as the stock of knowledge grows.

A fund manager's incentive to deliver performance declines as its stock of knowledge grows. Funds' AUM increases and decreasing returns to scale erode the risk-adjusted performance they deliver to investors. Yet, equilibrium AUM continues to rise because investors allocate more wealth when fund offerings align with their preferences, even absent alpha. This effect shifts the incentives for fund managers from generating alpha, to selling products that match investors' taste. This echoes [Ben-David et al. \(2023\)](#), who show that thematic ETFs deliver negative risk-adjusted performance after launch, on average.

1.8.1.3 Endogenous Choice of Adopting Data Technologies

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

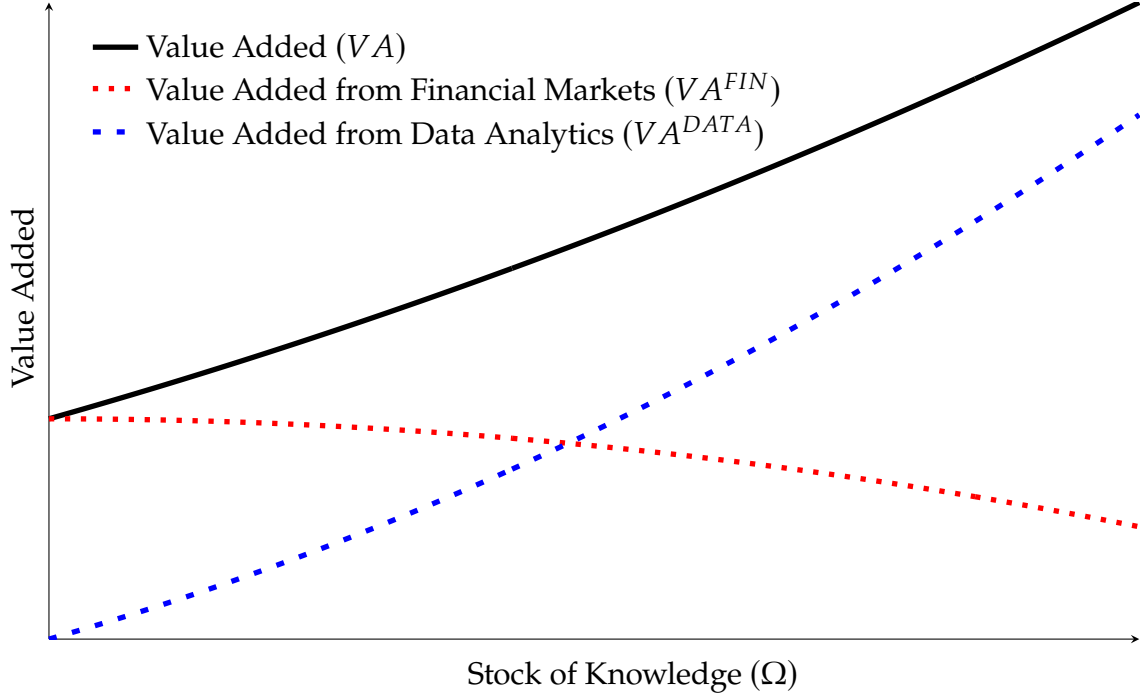


FIGURE 1.6: **Value Added and Stock of Knowledge.** The figure shows the value added a fund manager generates, as its stock of knowledge (data) grows. The black solid line depicts the total value added generated by an asset manager. The red dotted line shows the value added by extracting value from financial markets, while the blue dashed line represents the value added by analyzing data.

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

Investors' utility remains:

$$\text{TP}_{j,t} = a_j - b \cdot q_{i,t} - f_{i,t} + \kappa \cdot (\theta_t - \hat{\theta}_{j,t})^{-2}. \quad (1.22)$$

As before, asset managers filtering better investors' preference, θ_t , deliver higher utility all else equal. Managers who pay $\bar{c}$ observe more precise signals, match preferences more closely, and attract more capital. From equations (1.16) and (1.13), fund manager i 's optimal profits are

$$\pi_{j,t}^* = \frac{(a_j + \kappa \cdot \Omega_{i,t})^2}{4b} := \pi_k(a_j) \quad k \in \{D, T\}, \quad (1.23)$$

where $k \in \{D, T\}$ denotes whether the fund decides to adopt data analytics and being a "data-manager" (D), or not (T). In equilibrium, the benefit from adoption must

equal the cost $\bar{c}$. I first compute the gross profit gain from data analytics is:

$$\begin{aligned}\pi_D(a_j) - \bar{c} &= \frac{(a_j + \kappa \cdot \Omega_{j,t}^D)^2 - (a_j + \kappa \cdot \Omega_{j,t}^T)^2}{4b} = \\ &= \frac{\kappa^2[(\Omega_{j,t}^D)^2 - (\Omega_{j,t}^T)^2]}{4b} + \frac{\kappa(\Omega_{j,t}^D - \Omega_{j,t}^T) \cdot a_j}{2b}\end{aligned}\quad (1.24)$$

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

$$a^* = \underbrace{\frac{2b\bar{c}}{\Omega_{j,t}^D - \Omega_{j,t}^T}}_{\text{cost/benefit}} - \frac{\Omega_{j,t}^D + \Omega_{j,t}^T}{2}, \quad (1.25)$$

and a manager i adopts a technology to analyze customers' data if $a_j \geq a^*$, whereas when $a_j < a^*$ the manager does not adopt. The first term in equation (1.25) captures cost over benefit of adopting. If adoption is expensive (high $\bar{c}$) or the informational gain is small, no managers adopt. For instance, this may happen because it is very costly to acquire human capital trained in data analytics (high $\bar{c}$), or because the informational benefit of analyzing data is too little. On the other hand, when the cost-benefit term is small, many funds will install data technologies.

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

$$\lambda = \mathbf{P}\{a_j \geq a^*\} = 1 - \Phi\left(\frac{a^* - \bar{a}}{\tau_a^{-1}}\right). \quad (1.26)$$

In Section 1.4.3.2, I use the local supply of data analytics graduates to instrument adoption of a data technology. The intuition is that local experts reduce cost of adoption (first term in (1.25)), as it lowers the threshold a^* and raises the share of adopters.

1.8.2 Data Appendix

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

1.8.2.1 CRSP Mutual Funds

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

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

Following Berk and van Binsbergen (2015), I backfill missing CUSIP with the last available non-empty CUSIP within the same share class. This step replace 871,728 CUSIP obs. (9.55% of total). I do not *forward* fill missing observations. I replace missing *exp_ratio* and *actual_12b1* fees with their time series average within the same share class (Roussanov, Ruan and Wei, 2020). This step fills 646,928 obs. (7.09% of total) and 1,510,034 obs. (16.54% of total) respectively. Following Sirri and Tufano (1998) and Roussanov, Ruan and Wei (2020) I compute the "effective" 12b-1 fee summing CRSP's *actual_12b1* to the share class-month front load, and assuming the front load fee is amortized over 7 years.

I adjust AUM (TNA) for inflation to be comparable across time. The seasonally

adjusted monthly CPI is from FRED All Consumers: All Items, and I use January 2000 as baseline month (as in Berk and van Binsbergen, 2015).

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

1.8.2.2 FactSet Funds

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

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

1.8.2.3 Morningstar Direct

I complement CRSP Mutual Funds and FactSet data with information on US-domiciliated mutual funds from Morningstar Direct. I use monthly data at the share class level (*secid*). The raw Morningstar Direct dataset has 5,003,970 share class-month observations, and I will merge it with CRSP/FactSet data at the ticker-month level. Following Berk and van Binsbergen (2015), I start filling empty ticker observations with the last available non-empty ticker within share class. After this step, I set to “missing” all observations with more than one fund associated with the same ticker-month match, to avoid matching mistakes. Then, I drop all observations without a valid ticker (1,278,877 obs., 25.56 % of the initial raw data).

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

1.8.2.4 CRSP-Morningstar-FactSet Merged

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

The CRSP dataset resulting from Appendix Section 1.8.2.1 has 2,954 observations (0.03% of total) in which the same CUSIP-month pair appears twice. Inspecting those observations, they do not appear to be double reporting, but rather mistakes on CRSP's side. For each of those observations, I keep the CUSIP-month with the largest total AUM in the sample.

I merge CRSP and FactSet dataset by CUSIP-month. The merge results in 7,646,161/8,689,175 observations merged (88%).

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

I classify index funds following Berk and van Binsbergen (2015) and Pástor, Stambaugh and Taylor (2015) as closely as possible. I flag a share class observation as index fund if contains "INDEX", "ETF", "ISHARES", "IDX", "INDX" in its (uppercase letters) fund name, *et_flag* is either "F" or "N", the lipper class belongs to S&P500 index (*lipper_class* is "SPSP" or "SP"), the *index_fund_flag* is "Y", the FactSet's fund type is either "ETF" or "ETN", the (uppercase letters) brand name is "ISHARES", or it is an index levered fund. I identify a share classes to be of an index levered funds if it has FactSet's leverage factor larger than 1, or if it contains "INVERSE", "SHORT", "ULTRA", "2X", "3X", "4X", "5X", "6X", "7X", "8X", "9X", "0X", "SHORT TERM", "SHORT TM", "SHORT BOND", "SHORT BND", "LONG SHORT", "LG SHORT" in its (uppercase letters) fund name. I classify Vanguard funds if a fund's name matches one of these (uppercase letters) names: "VFINX", "VEXMX", "NAESX", "VEURX", "VPACX", "VVIAX", "VBINX", "VEIEX", "VIMSX", "VISGX", or "VISVX". Additionally, following Ben-David et al. (2018), I identify ETFs in my sample merging it with a list of ETFs from CRSP Monthly Stock File (*shrcd*==73). This merge results in 271,198/7,403,148 matches (3.66% of total.)

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

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

1.8.2.5 Website Technologies

I obtain information on a fund website's technology adoption from BuiltWith. In general, website technologies are defined as tools and services like analytics, payment systems, networking and programming scripts that enhance a website's features. For example, *PayPal Credit* is a technology that enables customers to make buy-now-pay-later payments on a website. Other examples of website technologies are *Google Maps*, *Apple Pay*, and *Shopify*. BuiltWith is a company specialized in website profiling, who sells these data to companies and consultants. They analyze websites' page code and search for specific patterns that identify the usage of technologies —similar to how a virus scanner searches for pattern in files to identify viruses. The most common patterns they use to identify such technologies are HTML tags, cookies, and Javascript snippets found in a website's source code. BuiltWith continuously crawls websites and analyzes their underlying technologies. They 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 consecutive crawls on a website's code. Appendix Figure 1.7 shows a snapshot of the technology data for *arrowfunds.com*, as it appears in my sample. For each unique website in my sample (from CRSP Mutual Funds data) I collect all the technologies name, and installation/removal dates. BuiltWith also provide a *technology_category* (e.g., Analytics, Feeds, etc.) for each technology, that I map to all technologies installed at least once in my sample. Then, I build a panel with *website-technology_name-month* where the first month is the *first_detected* month, and the last one is the *last_detected* month. I further filter for analytics' technologies that are aimed to collect and analyze customers' data, and I count the number of such technologies installed in the website. Finally, I merge this dataset by *website-date* to the main CRSP/Morningstar/FactSet data at the share class level (before aggregation). I merge 3,091,431/7,614,808 observations (i.e., 40.60% of the share class-month observations

starting March 1993 have at least one data technology in place). Then, I replace missing with zeros, if I have a valid website for the share class-month observation and I have data from BuiltWith for the associated *website*-month, but BuiltWith does not detect data analytics technologies for that month.

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

FIGURE 1.7: Example of website technology data. This figure displays a snapshot of the website technology data from BuiltWith for arrowfunds.com (in my sample).

1.8.2.6 Final Funds Sample

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

$$Flow_{i,t} = \frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t}) - MGN_{i,t}}{AUM_{i,t-1}}, \quad (1.27)$$

where $AUM_{i,t}$ represents total net assets for fund i in month t , $r_{i,t}$ is the (gross) monthly return, and $MGN_{i,t}$ is the increase in AUM due to the fund's mergers (if any) in month t . Since CRSP does not reports the exact date in which the merger takes place, I follow Lou (2012) and use information about the latest available NAV of the target funds to build a six-months window where the merger plausibly took place. In particular, from CRSP I observe the last date in which the target fund has non-empty NAV, and the identifier of the acquirer. For the acquiring fund, I build a six months window which starts one month before the latest available date of the target fund, until five months after. Within this window, I first compute the flows without accounting for the possible merger (i.e., $\frac{AUM_{i,t} - AUM_{i,t-1} \cdot (1 + r_{i,t})}{AUM_{i,t-1}}$) and I flag as *merger_month* the date with highest flow within the six-months window. Appendix Table 1.16 gives an example of this approach for the acquirer fund *crsp_fundno* == 662. In the example, the target fund had latest AUM of \$452.5 million (latest date 1999m4). Around the six-months window, the acquirer has one clear *flow* outlier (computed without accounting for the merger); i.e., a +2,294% net flow in 1995m5. I flag the 1999m5 observation as merger date. Therefore, following equation 1.27, the actual fund flow in 1999m5 is -0.6417.

<i>crsp_fundno</i>	<i>month</i>	<i>TNA</i>	<i>window6M</i>	<i>trgt_lastTNA</i>	<i>flow</i>	<i>flagMerger</i>
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 1.16: **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).

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

1.8.2.7 N-SAR Filings

Form N-SAR filings are SEC regulatory files that US registered investment companies (including open-end mutual funds in my sample) used to report until June 1, 2018.³⁸ N-SAR filings were filed semi-annually and contains a host of detailed information including a fund's service providers, distribution activity, brokerage, and governance. I obtain N-SAR regulatory filings data from Wharton Research Data Service (WRDS) from January 2006, to June 2018. Importantly, each "registrant" (or fund complex) files a separate N-SAR form and reports information for all funds within the fund complex. Each registrant has a unique *cik* code, but it can include more than one fund. To extract fund-level information in a registrant group, I use three approaches. First, for registrant groups with only one fund the approach is straightforward: I link the information in the Form N-SAR to the (unique) fund reported by the registrant. This step yields 18,843 observations.

Second, for registrants with more than one fund, I first match the registrant group with all funds contained in the complex, by *cik*-month. Then, I fuzzy match by funds' name (*series_name*) within registrant group using Stata command `RECLINK2`. I require that the precision of the fuzzy match is above 97.5%. This procedure gives 185,236 perfect matches, and 75,688 fuzzy matches with average precision 99.80%.

Third, for funds without a valid *series_name*, I match to the respective registrant by *cik* code. Then, for all funds in a fund complex I obtain the turnover and net flows in the N-SAR filings and compute all the pairwise correlations with turnover and net flows from the main CRSP/Morningstar/FactSet dataset (see Appendix 1.8.2.4) within *cik*. Finally, I link funds with pairwise correlation of turnover *and* net flows above 99%. After this step, I have 595 additional observations.

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

1.8.2.8 Mutual Fund Prospectus Data

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

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

First, I exclude spurious matches near the table of contents or summary sections. These PIS candidate are typically short texts (between 5 and 20 words) around the summary section of the fund's prospectus, and clearly represent false positive PIS sections. Second, I require that the section contain characteristic language associated with investment strategies, such as "under normal circumstances...", "the fund seeks..." or "the fund normally invests primarily...". I clean each section by removing formatting characters, and stripping out numbers unlikely to represent percentages or portfolio weights. Appendix A in (Abis and Lines, 2024, p. 21) shows that the cross sectional distribution of PIS word counts ranges from 20 to around 1,500. Therefore, I remove candidate PIS section with less than 20 words, or excessively long blocks (>2,500 words).

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

1.8.2.9 Mutual Funds Portfolio Holdings

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

Previous research (e.g., Shive and Yun, 2013) reports a discontinuity in Thomson Reuters (s12) mutual funds holdings coverage after 2008, when compared to CRSP holdings. A common solution in the literature has been to use s12 holdings before June 2008 and CRSP thereafter. However, Thomson Reuters seems to have solved this issue in later vintages (see Appendix Figure 1.8). Thus, I prefer not to append different data sources and use the updated s12 holdings throughout the sample.

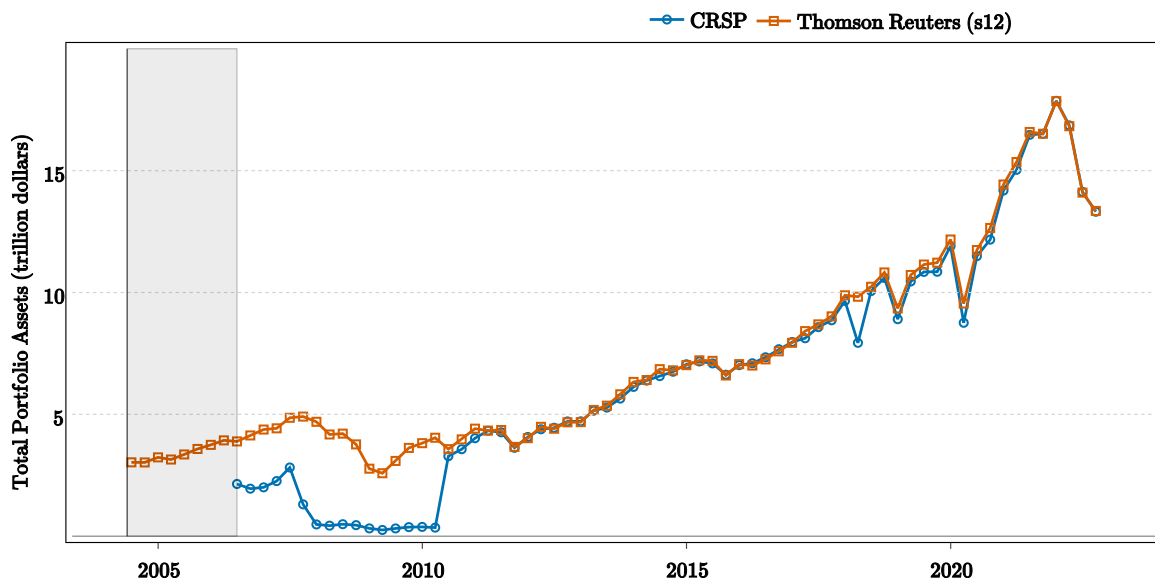


FIGURE 1.8: **Total Assets in Thomson Reuters (s12) and CRSP Holdings.** The figure shows the total assets (in trillion of US dollars) in the Thomson Reuters (s12) and CRSP mutual funds holdings databases. CRSP has no holdings data before 2003 (shaded area). The solid blue line reports total equity assets 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.

1.8.3 List of CIP Codes for Instrumental Variable

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

TABLE 1.17: **CIP Codes for Data Analytics-Related Fields:** This table reports the Core Instructional Programs (CIP) that I use to construct the local supply of graduates in data analytics. I obtain degree counts from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees awarded by U.S. universities. The annual sample period is from 2000 to 2023.

1.8.4 Mutual Funds Prospectuses Text Classification

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

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

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

Thematic categories proxy for a fund tends to mention retail-oriented themes in its PIS. I classify words and phrases into themes such as ESG, AI, or clean energy. I select themes based on examples in (Ben-David et al., 2023). Table 1.18 provides the full list of tokens and phrases by theme. I compute the composite measure as the sum of all themes mentions.

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

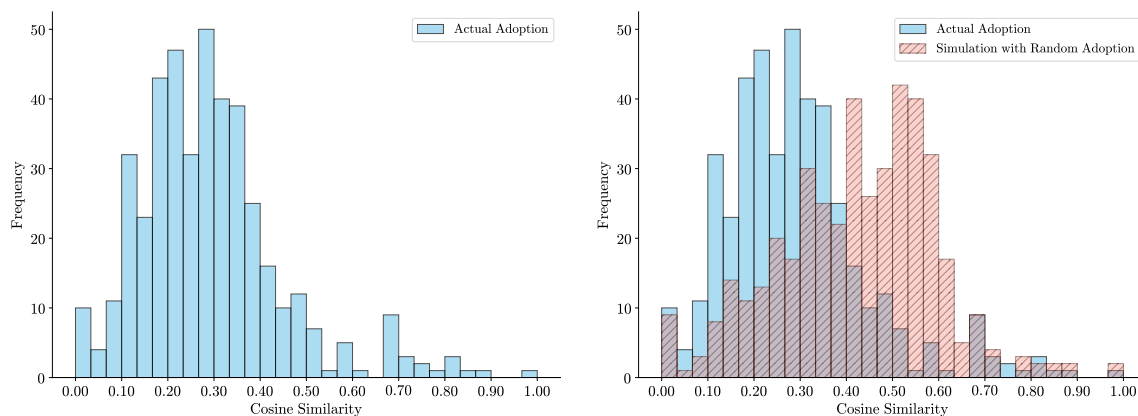
Category	Tokens and Phrases
ESG	esg, sustainable, impact, environmental, governance, responsible
Clean Energy	renewable, solar, wind, battery, ev, "clean energy", "energy transition"
AI	ai, robotics, neural, "artificial intelligence", "machine learning", "generative ai"
Cryptocurrency	crypto, bitcoin,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

TABLE 1.18: **Themes Classification and Words/Phrases:** This table reports the list of words and sentences associated with each retail-oriented theme. I count the number of word phrase matches in each fund PIS, scale by the total number of words, and express the final measure per 100 words of text.

1.8.5 Additional Results

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

1.8.5.1 Appendix Figures



(A:) Overlap in Technology Adoption.

(B:) Comparison with Random Simulation.

FIGURE 1.9: Overlap in Technology Adoption within Hosting Platform. The figure shows the cosine similarity in website technology adoption within website hosting platforms. Cosine similarity is higher when there is high overlap among website technologies within a given hosting platform. Panel A shows the histogram of similarities for the sample of US mutual funds in this paper (mean 0.29, median 0.27). Panel B also reports, in red, the histogram of similarities for counterfactual simulations where technology adoption within hosting platforms is random (mean 0.43, median 0.44).

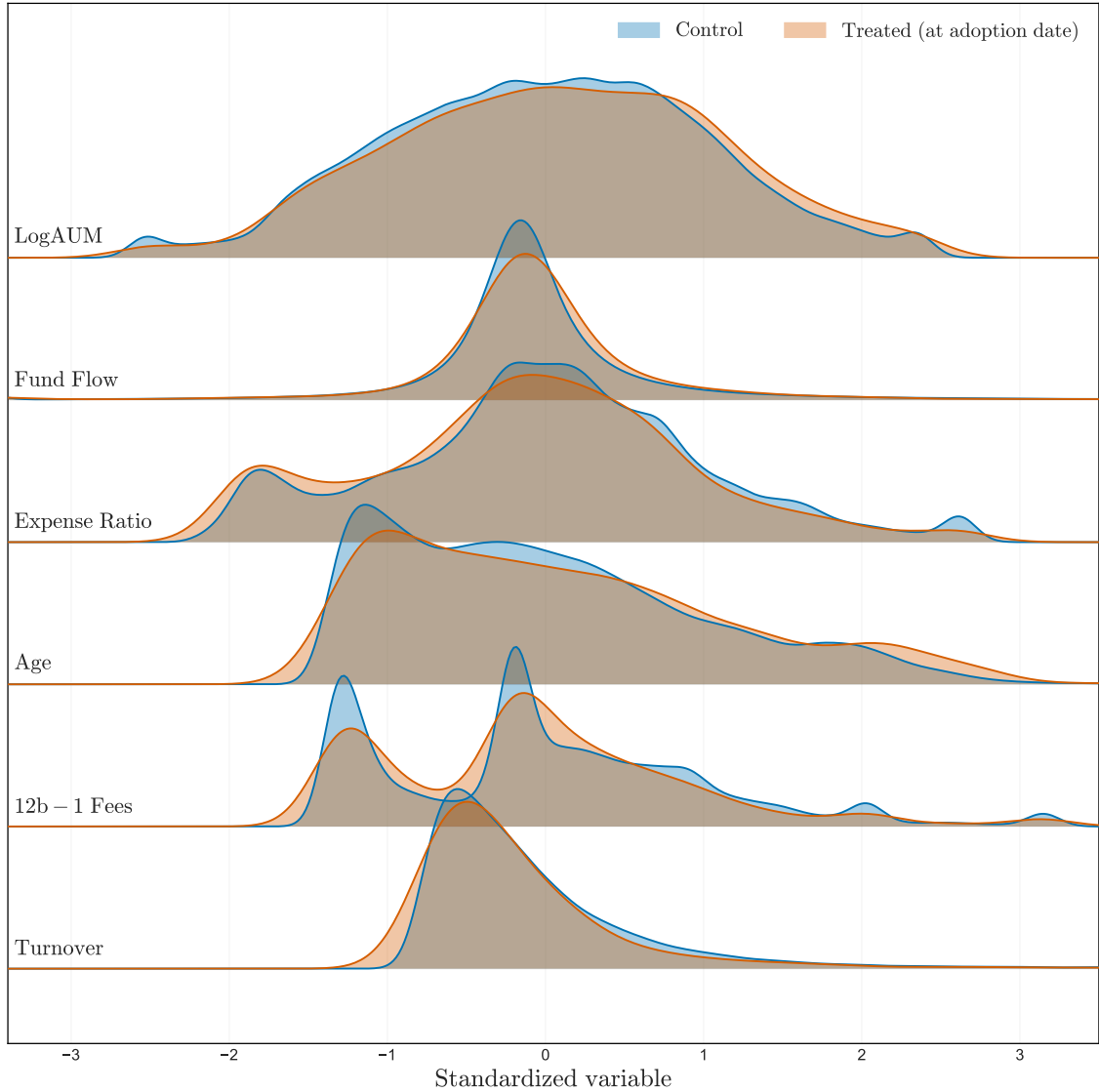


FIGURE 1.11: Balance Covariates. This figure shows the empirical pdf of covariates for funds in the treatment and control group. All variables are normalized to have a mean of zero and a standard deviation of one in the full sample. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t . For each group, I report the covariates in the 3 months pre-adoption. The figure shows covariates included in regressions in the main text.

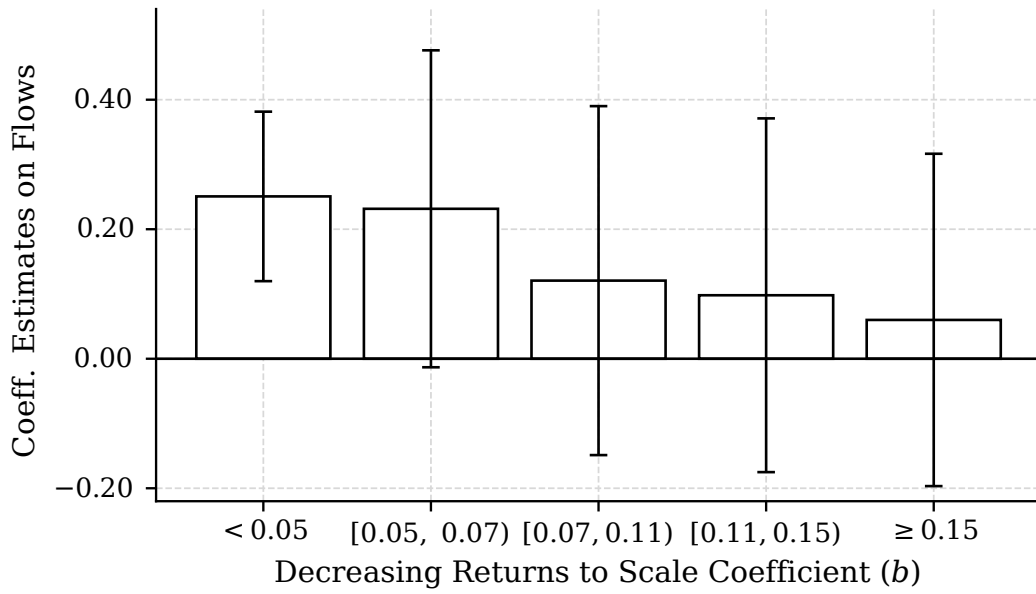


FIGURE 1.12: Heterogeneity of Effect on Flows by Decreasing Returns to Scale. This figure shows results for difference-in-differences coefficients across different values of decreasing returns to scale (b) as of data technology adoption. I estimate decreasing returns to scale as-of adoption following the recursive demeaning approach proposed in [Pástor, Stambaugh and Taylor \(2015\)](#). The coefficient b captures the extent to which a fund's size reduces its ability to generate alpha. Each bar represents a level of subsample by decreasing returns to scale coefficient (e.g., the first bar represents all funds with almost negligible decreasing returns to scale). Each vertical line represents the 95% confidence interval. The specification is the same as the main specification in equation (1.2). All regressions include fund and category×time fixed effects, and controls: fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t .

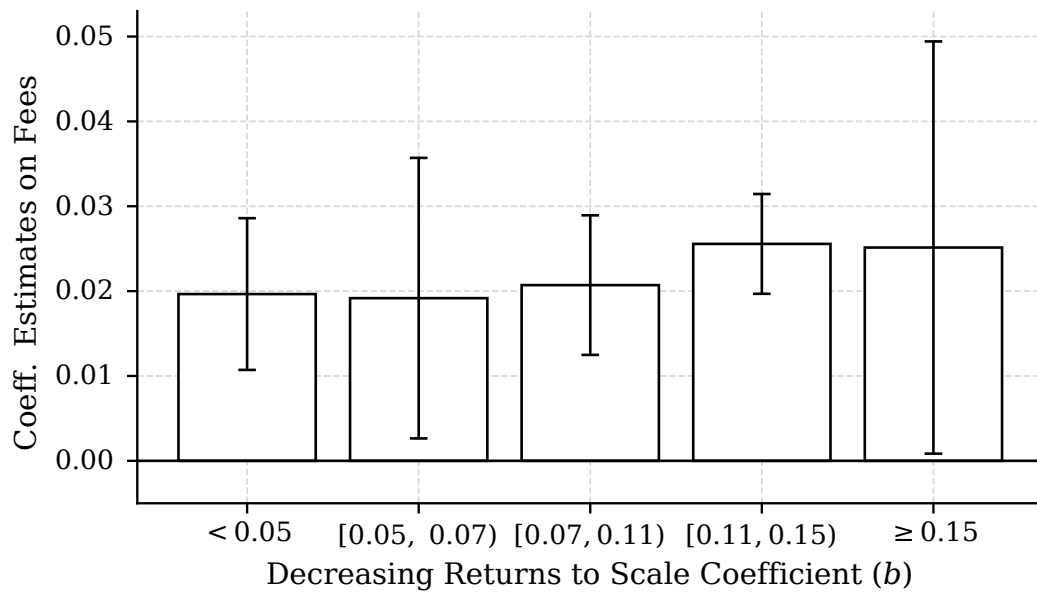


FIGURE 1.13: **Heterogeneity of Effect on Expense Ratio by Decreasing Returns to Scale.** This plot report a placebo test on the heterogeneity of the effect of data technology on expense ratio by subsample based on decreasing returns to scale. The figure shows results for difference-in-differences coefficients across different values of decreasing returns to scale (b) as of data technology adoption. I estimate decreasing returns to scale as-of adoption following the recursive demeaning approach proposed in [Pástor, Stambaugh and Taylor \(2015\)](#). The coefficient b captures the extent to which a fund's size reduces its ability to generate alpha. Each bar represents a level of subsample by decreasing returns to scale coefficient (e.g., the first bar represents all funds with almost negligible decreasing returns to scale). Each vertical line represents the 95% confidence interval. The specification is the same as for Table 1.5. All regressions include fund and category×time fixed effects, and controls: fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t .

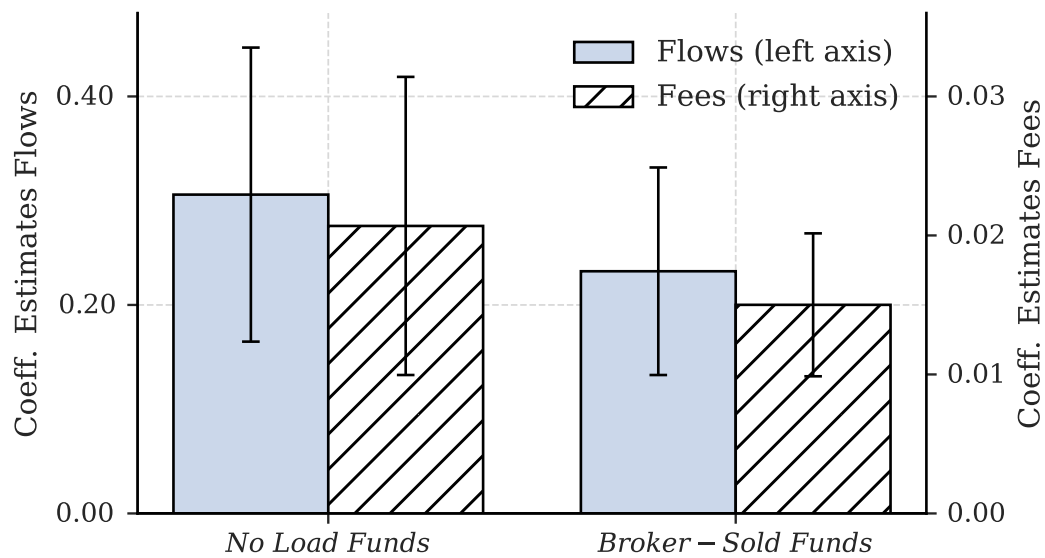


FIGURE 1.14: Direct Sold vs Broker-Sold Funds. This figure shows results for difference-in-differences coefficients for no-load funds and broker-sold funds. No-load funds are funds that are directly sold to investors [Del Guercio and Reuter \(2014\)](#). I identify directly sold (no-load) funds and broker-sold funds following [\(Sun, 2021\)](#). Each bar represents a coefficient estimates on fund flows (plain bars) or expense ratio (hatched bars). The left (right) axis refers to flows (fees). Vertical line represents the 95% confidence interval. The specification is the same as the main specification in equation (1.2). All regressions include fund and category×time fixed effects, and controls: fund’s size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t .

1.8.5.2 Appendix Tables

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

TABLE 1.19: **Fund Flows and Data Technologies, detailed covariates results:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($DATA_{i,t}$), and controls for fund-month characteristics. See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.20: **Parallel trends:** This table reports the growth rate of fund flows in the 12 months pre-adoption, for adopting and not-adopting fund. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t . The table shows monthly and quarterly growth rates of fund flows for the sample of treated and control group in the 12 months pre-adoption. I winsorize growth rates at the 1% and 99% level. The last column reports the p-value of the difference between treated and control groups.

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

TABLE 1.21: **Covariates balance.:** This table reports summary statistics of covariates for funds in the treatment and control group. For each fund's adoption of a data technology in month t (treatment), I construct the control group as the sample of funds with no data technologies in place in month t . For each group, I report the covariates in the 3 months pre-adoption. The table shows covariates included in regressions in the main text. AUM is inflation adjusted in January 2000 \$ million. Expense Ratio, 12b-1 Fees, and Flows are in %; e.g., the average fund flow for the control group is 0.20% monthly.

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

TABLE 1.22: **Fund Flows and Data Technologies, additional controls:** This table shows results of panel regression robust to several additional controls in addition to the baseline regression. The dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . In columns (1), (2), (5), and (6) I also include past fund flow and expense ratio as controls. Columns (3), (4), (7), and (8) add the Morningstar Rating in the set of controls. Columns (1) to (4) report results for baseline OLS, while columns (5) to (8) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.23: **Fund Size and Data Technologies:** This table shows results of panel regression in which the dependent variable is the fund size –that is, a fund’s ($\log$) AUM. The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($\text{DATA}_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund’s ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . I exclude the lagged fund’s size as a control, because it might introduce post-treatment bias (Roberts and Whited, 2013). Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.24: **Expense Ratio and Data Technologies, additional controls:** This table shows results of panel regression robust to several additional controls in addition to the baseline regression. The dependent variable is the one-month-ahead fund expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA _{i,t}), and controls for fund-month characteristics. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . In columns (1), (2), (5), and (6) I also include past fund flow and expense ratio as controls. Columns (3), (4), (7), and (8) add the Morningstar Rating in the set of controls. Columns (1) to (4) report results for baseline OLS, while columns (5) to (8) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.25: **Fund Flows and Data Technologies using different measures of performance:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The baseline control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees. Columns (1) and (2) additionally control for alpha with respect to a set of passive Vanguard funds, following Berk and van Binsbergen (2015). Columns (3) and (4) use Fama and French (1993) 3-Factors, while columns (5) and (6) use alpha with respect to the Fama and French (2015) 3-Factors model. All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.26: **Expense Ratio and Data Technologies using different measures of performance:** This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA _{i,t}), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The baseline control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees. Columns (1) and (2) additionally control for alpha with respect to a set of passive Vanguard funds, following Berk and van Binsbergen (2015). Columns (3) and (4) use Fama and French (1993) 3-Factors, while columns (5) and (6) use alpha with respect to the Fama and French (2015) 3-Factors model. All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.27: **Robustness of Main Results excluding Google Analytics:** This table addresses concerns that the results are entirely driven by Google Analytics. The table replicates the main findings in Tables 1.3 and 1.5 excluding Google Analytics from the set of data technologies. Columns (1) and (2) show results for flows, while columns (3) and (4) report the robustness for results on expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . All estimates use difference-in-differences estimator robust to staggered treatment design (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.28: **Robustness Results on Flows, clustering by Fund Family:** This table shows robustness results for the main specification, clustering standard errors by fund family. The dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.29: **Robustness Results on Expense Ratio, clustering by Fund Family:** This table shows robustness results for the main specification, clustering standard errors by fund family. The dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.30: **Robustness for Fund Families with only one website for each fund:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow, for a subset of the main sample. I keep only fund families with a unique website associated to a given fund, and run the main specification. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA $_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.31: **Robustness aggregating Flows by Fund Family:** This table shows results of panel regression in which the dependent variable is the one-month-ahead within fund family total flow. The regressors are a dummy equal to one if a fund family i has at least a data technology in place at month t (DATA $_{i,t}$), and controls for fund family-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund family's total size ($\log$ total AUM), ($\log$) age, and the number of funds in the family in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.32: **Robustness Results on Expense Ratio, Retail Share Classes Only:** This table shows robustness results for the main specification on expense ratio, for retail share classes only. For each fund, I aggregate only retail share classes and remove all institutional share classes. Then, I run the main specification as in Table 1.5. The dependent variable is the one-month-ahead expense ratio. The regressors are a dummy equal to one if a fund i has a data technology in place at month t (DATA_{*i,t*}), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . Columns (1) and (2) report results for baseline OLS, while columns (3) and (4) show results using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund family and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.33: **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 ($\log \text{AUM}$) and ($\log$) age. All standard errors are clustered by month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.34: **Robustness IV Estimates on Flows: Excluding Financial Districts:** This table shows robustness results for the instrumental variable estimates where the instrument is the local supply of talent in data analytics-related fields. The dependent variable is the one-month-ahead fund flow. In column (1) to (6), I remove funds located in each financial district as defined in [Christoffersen and Sarkissian \(2009\)](#), namely, Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y , with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix 1.8.3 for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. All estimates use difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y ($DATA_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in year y . All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.35: **Robustness IV Estimates on Fees: Excluding Financial Districts:** This table shows robustness results for the instrumental variable estimates where the instrument is the local supply of talent in data analytics-related fields. The dependent variable is the one-month-ahead expense ratio. In column (1) to (6), I remove funds located in each financial district as defined in [Christoffersen and Sarkissian \(2009\)](#), namely, Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco. I instrument the adoption choice of fund i in commuting zone (CBSA) c and year y , with the number of graduates in data analytics, statistics, and computer science from universities within a fund's CBSA. Annual university graduates are from the Integrated Postsecondary Education Data System (IPEDS) and include bachelor's, master's, and Ph.D. degrees. See Appendix 1.8.3 for the detailed list of Core Instructional Programs (CIPs) in data analytics, statistics, and computer science. All estimates use difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). The regressors are a dummy equal to one if a fund i in commuting zone c has a data technology in place in year y ($DATA_{i,c,y}$), and controls for fund-year characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in year y . All variables are at the annual frequency. The sample period is from 2000 to 2023. All standard errors are two-way clustered by fund and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.36: **Fund Fees and TensorFlow:** This table shows results of panel regression in which the dependent variable is the one-month-ahead expense ratio. Columns (1) and (2) follow specification in equation (1.3), while columns (3) to (6) follow (1.4). In columns (3) and (4) the continuous treatment z_i is the (*log*) number of months between the first data technology adoption and TensorFlow’s release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow’s release, as continuous treatment z_i . DATA_{*i,t*} is a dummy equal to one if fund *i* has a data technology in place at month *t*. See Section 1.3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund’s size (*log*AUM), (*log*) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (1.6)) in month *t*. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow’s release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	Fund Flows $_{i,t+1}$ (%)					
			z_i : Tenure of Adoption		N. of Data Tech.	
	(1)	(2)	(3)	(4)	(5)	(6)
$DATA_{i,t}$	0.567*** (0.139)	0.595*** (0.140)	0.517*** (0.146)	0.517*** (0.146)	0.491*** (0.136)	0.491*** (0.136)
$DATA_{i,t} \times Post_t$	0.253** (0.109)	0.313*** (0.109)				
$DATA_{i,t} \times Post_t \times z_i$			0.080** (0.033)	0.080** (0.033)	0.023** (0.011)	0.023** (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 1.37: **Robustness on TensorFlow, Excluding Growth Funds:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. I exclude growth funds from this sample to mitigate concerns that results are driven by an increase in expected cash flows for tech stocks. Columns (1) and (2) follow specification in equation (1.3), while columns (3) to (6) follow (1.4). In columns (3) and (4) the continuous treatment z_i is the ($\log$) number of months between the first data technology adoption and TensorFlow's release. Columns (5) and (6) use the number of data technologies installed as of TensorFlow's release, as continuous treatment z_i . $DATA_{i,t}$ is a dummy equal to one if fund i has a data technology in place at month t . See Section 1.3.2 for details on data technologies. The fund-month control variables (omitted for brevity) include a fund's size ($\log AUM$), ($\log$) age, turnover, CAPM alpha, 12b-1 fees, and the coefficient of data competition (equation (1.6)) in month t . The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023, which did not adopt a data technology after June 2015 (i.e., six-months before TensorFlow's release). All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.38: **Fund Flows and Competition:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($\text{DATA}_{i,t}$), and controls for fund-month characteristics. See Section 1.3.2 for details on data technologies. I include an interaction term with the competition coefficient $\gamma_{c,t}$ for fund category c in month t . The competition coefficient is built following equation (1.6), and it captures the fraction of funds with data technologies in place within fund category-month. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.39: **Fund Flows and Number of Technologies:** This table shows results of panel regression in which the dependent variable is the one-month-ahead fund flow. The regressors are the number of data technologies in place for fund *i* at month *t* (N. *Tech*_{*i,t*}) in column (1); column (2) adds its square (N. *Tech*_{*i,t*}²), and the *log* of (1+N. *Tech*_{*i,t*}) in columns (3)-(4). See Section 1.3.2 for details on data technologies. All columns include controls for fund-month characteristics (omitted for brevity). The control variables are a fund's size (*logAUM*), (*log*) age, turnover, and CAPM alpha in month *t*. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.40: **Performance and Data Technologies:** This table shows results of regression in which the dependent variable is the one-month-ahead fund performance. Columns (1), (3), (5), and (7) report results for OLS regressions, while columns (2), (4), (6) and (8) use the recursive demeaning approach (RD) in [Pástor, Stambaugh and Taylor \(2015\)](#) which accounts for the positive contemporaneous correlation between fund size and unexpected returns. The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($DATA_{i,t}$), 12b-1 fees, and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, and turnover. I use Vanguard alpha (columns (1) and (2)), CAPM alpha (columns (3) and (4)), FF-3 factors alpha (columns (5) and (6)), and FF-5 factors alpha (columns (7) and (8)) as proxy of funds' performance. The Vanguard alpha is the risk adjusted performance with respect to the set of available Vanguard index funds ([Berk and van Binsbergen, 2015](#)). The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.41: **Redistribution of 12b-1 Fees Payments:** This table shows results of panel regression on funds' distribution to 12b-1 payment categories (in %), robust to concerns in staggered difference-in-differences (see [Goodman-Bacon, 2021](#)). The dependent variable is the fund's one-month-ahead percentage allocation of 12b-1 fees in each category. Column (1) report results for the percentage payment towards broker and dealers, column (2) to captive retail sales force, column (3) to the underwriter itself (retained), column (4) to external advertising, and column (5) to printing and mailing of prospectuses to prospect clients. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). The regressors are a dummy equal to one if a fund i has a data technology in place at month t ($DATA_{i,t}$), and controls for fund-month characteristics (omitted for brevity). See Section 1.3.2 for details on data technologies. The control variables include a fund's size ($\log AUM$), ($\log$) age, turnover, 12b-1 fees, and CAPM alpha in month t . The monthly sample is from January 2006 to June 2018. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

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

TABLE 1.42: **Flow-Performance Sensitivity and Data Technologies:** This table shows results of panel regression of flow-performance sensitivity and interaction terms with the adoption of a data technology. The dependent variable is the one-month-ahead fund flow. The regressors are the fund's performance in month t , and interaction term with a dummy equal to one if a fund i has a data technology in place at month t (DATA_{*i,t*}), and controls for fund-month characteristics. The control variables include a fund's size ($\log$ AUM), ($\log$) age, turnover, and 12b-1 fees in month t . The first row reports results for the flow-performance sensitivity (Chevalier and Ellison, 1997). The second row shows results for the interaction term of flow-performance sensitivity with the DATA dummy. Columns (1)-(2) use CAPM alpha as measure of a fund's performance, columns (3)-(4) use the alpha with respect to a set of 11 existing Vanguard funds (Berk and van Binsbergen, 2015), columns (5)-(6) use Fama-French 3-factors alpha, and columns (7)-(8) use Fama-French 5-factors alpha. The monthly sample include equity mutual funds and ETFs from March 1993 to December 2023. All standard errors are two-way clustered by fund and month (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

References

- Abis, Simona (2022) "Man vs. Machine: Quantitative and Discretionary Equity Management," *Working paper*.
- Abis, Simona and Anton Lines (2024) "Broken promises, competition, and capital allocation in the mutual fund industry," *Journal of Financial Economics*, Vol. 162, p. 103948.
- Abis, Simona and Laura Veldkamp (2023) "The Changing Economics of Knowledge Production," *The Review of Financial Studies*, Vol. 37, No. 1, pp. 89–118, 08.
- Amazon (2025) "Amazon Cookie Policy," Technical report.
- Arellano-Bover, Jaime, Carolina Bussotti, Matteo Paradisi, and Liangjie Wu (2025) "The Labor Demand Implications of Brand Capital: Insights from Trademark Transactions in Italy," *Working paper*.
- Argyle, Bronson, Taylor Nadauld, and Christopher Palmer (2022) "Real Effects of Search Frictions in Consumer Credit Markets," *The Review of Financial Studies*, Vol. 36, No. 7, pp. 2685–2720, 11.
- Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024) "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, Vol. 151.
- Baker, Scott R., Brian Baugh, and Marco Sammon (2023) "Customer Churn and Intangible Capital," *Journal of Political Economy Macroeconomics*, Vol. 1, No. 3, pp. 447–505.
- Baley, Isaac and Laura L. Veldkamp (2025) *The Data Economy: Tools and Applications*, Princeton, NJ: Princeton University Press.
- Barber, Brad M., Xing Huang, and Terrance Odean (2016) "Which Factors Matter to Investors? Evidence from Mutual Fund Flows," *The Review of Financial Studies*, Vol. 29, No. 10, pp. 2600–2642, October.
- Basten, Christoph and Steven Ongena (2020) "The Geography of Mortgage Lending in Times of FinTech," *CEPR Discussion Paper*, No. DP14918.
- Belo, Frederico, Xiaoji Lin, and Maria Ana Vitorino (2014) "Brand capital and firm value," *Review of Economic Dynamics*, Vol. 17, No. 1, pp. 150–169.
- Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi (2018) "Do ETFs Increase Volatility?," *The Journal of Finance*, Vol. 73, No. 6, pp. 2471–2535.
- Ben-David, Itzhak, Francesco Franzoni, Byungwook Kim, and Rabih Moussawi (2023) "Competition for Attention in the ETF Space," *The Review of Financial Studies*, Vol. 36, No. 3, pp. 987–1042, 08.
- Ben-David, Itzhak I., Jiacui Li, Andrea Rossi, and Yang Song (2022) "What Do Mutual Fund Investors Really Care About?" *The Review of Financial Studies*, Vol. 35, No. 4, pp. 1723–1774, April.
- Berk, Jonathan B. and Richard C. Green (2004) "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, Vol. 112, No. 6, pp. 1269–1295.

- Berk, Jonathan B. and Jules H. van Binsbergen (2015) "Measuring skill in the mutual fund industry," *Journal of Financial Economics*, Vol. 118, No. 1, pp. 1–20.
- (2016) "Assessing asset pricing models using revealed preference," *Journal of Financial Economics*, Vol. 119, No. 1, pp. 1–23.
- Betermier, Sebastien, David Schumacher, and Ali Shahradeh (2023) "Mutual Fund Proliferation and Entry Deterrence," *The Review of Asset Pricing Studies*, Vol. 13, No. 4, pp. 784–829.
- Birru, Justin, Sinan Gokkaya, Xi Liu, and Stanimir Markov (2024) "Quants and market anomalies," *Journal of Accounting and Economics*, Vol. 78, No. 1.
- Bonelli, Maxime (2024) "Data-driven Investors," *Working paper*.
- Bonelli, Maxime and Thierry Foucault (2024) "Displaced by Big Data: Evidence from Active Fund Managers," *Working paper*.
- Bonelli, Maxime, Anastasia Buyalskaya, and Tianhao Yao (2023) "Information Intermediaries and Financial Product Design: Evidence from Mutual Funds," *Working paper*.
- Broadridge (2023) "Market Analytics," *Technical report*.
- Brynjolfsson, Erik and Kristina McElheran (2016) "The Rapid Adoption of Data-Driven Decision-Making," *American Economic Review*, Vol. 106, No. 5, p. 133–39, May.
- Buchak, Greg, Vera Chau, and Adam Jørring (2023) "Integrated Intermediation and Fintech Market Power," *Swiss Finance Institute Research Paper No. 23-67*.
- Callaway, Brantly and Pedro H.C. Sant'Anna (2021) "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, Vol. 225, No. 2, pp. 200–230.
- Carlin, Bruce I. (2009) "Strategic price complexity in retail financial markets," *Journal of Financial Economics*, Vol. 91, No. 3, pp. 278–287.
- Cen, Xiao, Winston Wei Dou, Leonid Kogan, and Wei Wu (2024) "Fund Flows and Income Risk of Fund Managers," *Working paper*.
- Charoenwong, Ben, Zachary T. Kowaleski, Alan Kwan, and Andrew G. Sutherland (2024) "RegTech: Technology-driven compliance and its effects on profitability, operations, and market structure," *Journal of Financial Economics*, Vol. 154.
- Chen, Jiafeng and Jonathan Roth (2023) "Logs with Zeros? Some Problems and Solutions*," *The Quarterly Journal of Economics*, Vol. 139, No. 2, pp. 891–936, 12.
- Chen, Jane, Wenxi Jian, and Mindy Z. Xiaolan (2022) "The Economics of Mutual Fund Marketing," *Working paper*.
- Chernenko, Sergey and Adi Sunderam (2020) "Do fire sales create externalities?" *Journal of Financial Economics*, Vol. 135, No. 3, pp. 602–628.
- Chevalier, Judith and Glenn Ellison (1997) "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, Vol. 105, No. 6, pp. 1167–1200.
- Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng (2024) "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts," *Management Science*.

- Christoffersen, Susan E.K. and Sergei Sarkissian (2009) "City size and fund performance," *Journal of Financial Economics*, Vol. 92, No. 2, pp. 252–275.
- Christoffersen, Susan, David K. Musto, and Russell Wermers (2014) "Investor Flows to Asset Managers: Causes and Consequences," *Annual Review of Financial Economics*, Vol. 6, No. 1, pp. 289–310.
- Chung, Cindy and Laura Veldkamp (2024) "Data and the Aggregate Economy," *Journal of Economic Literature*, Vol. 62, No. 2, p. 458–84, June.
- Coleman, Braiden, Kenneth Merkley, and Joseph Pacelli (2022) "Human Versus Machine: A Comparison of Robo-Analyst and Traditional Research Analyst Investment Recommendations," *The Accounting Review*, Vol. 97, No. 5, pp. 221–244.
- Cong, Lin W., Danxia Xie, and Longtian Zhang (2021) "Knowledge Accumulation, Privacy, and Growth in a Data Economy."
- Conley, Timothy G. and Christopher R. Udry (2010) "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, Vol. 100, No. 1, p. 35–69.
- Cremers, K.J. Martijn and Antti Petajisto (2009) "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *The Review of Financial Studies*, Vol. 22, No. 9, pp. 3329–3365.
- Cruz, Ryan E., James M. Leonhardt, and Todd Pezzuti (2017) "Second Person Pronouns Enhance Consumer Involvement and Brand Attitude," *Journal of Interactive Marketing*, Vol. 39, pp. 104–116.
- Cujean, Julien (2020) "Idea sharing and the performance of mutual funds," *Journal of Financial Economics*, Vol. 135, No. 1, pp. 88–119.
- Cvitanić, Jakša and Julien Hugonnier (2022) "Optimal fund menus," *Mathematical Finance*, Vol. 32, No. 2, pp. 455–516.
- Célérier, Claire and Boris Vallée (2017) "Catering to Investors Through Security Design: Headline Rate and Complexity," *The Quarterly Journal of Economics*, Vol. 132, No. 3, pp. 1468–1508.
- D’Acunto, Francesco and Alberto G. Rossi (2023) "Robo-Advice: Transforming Households into Rational Economic Agents," *Annual Review of Financial Economics*, Vol. 15, No. Volume 15, 2023, pp. 543–563.
- Dannhauser, Caitlin D. and Harold D. Spilker (2023) "The Modern Mutual Fund Family," *Journal of Financial Economics*, Vol. 148, No. 1, pp. 1–20.
- de Chaisemartin, Clément and Xavier D’Haultfoeuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, Vol. 110, No. 9, p. 2964–96, September.
- deHaan, Ed, Yang Song, Chloe Xie, and Christina Zhu (2021) "Obfuscation in mutual funds," *Journal of Accounting and Economics*, Vol. 72, No. 2.
- Del Guercio, Diane and Jonathan Reuter (2014) "Mutual Fund Performance and the Incentive to Generate Alpha," *The Journal of Finance*, Vol. 69, No. 4, pp. 1673–1704.

- Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2024) "Does Alternative Data Improve Financial Forecasting? The Horizon Effect," *The Journal of Finance*, Vol. 79, No. 3, pp. 2237–2287.
- Dou, Winston Wei, Leonid Kogan, and Wei Wu (2024) "Common Fund Flows: Flow Hedging and Factor Pricing," *The Journal of Finance* (forthcoming).
- Dugast, Jerome and Thierry Foucault (2024) "Equilibrium Data Mining and Data Abundance," *The Journal of Finance*.
- Edelen, Roger M. (1999) "Investor flows and the assessed performance of open-end mutual funds," *Journal of Financial Economics*, Vol. 53, No. 3, pp. 439–466.
- Ellison, Glenn and Sara Fisher Ellison (2009) "Search, Obfuscation, and Price Elasticities on the Internet," *Econometrica*, Vol. 77, No. 2, pp. 427–452.
- Evans, Richard B. (2010) "Mutual Fund Incubation," *The Journal of Finance*, Vol. 65, No. 4, pp. 1581–1611.
- Evans, Richard, Juan-Pedro Gomez, and Rafael Zambrana (2024) "MiFID II Research Unbundling: Cross-border Impact on Asset Managers," *Working paper*.
- Fama, Eugene F. and Kenneth R. French (1993) "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, Vol. 33, No. 1, pp. 3–56.
- Fama, Eugene F. and Kenneth R. French (2015) "A five-factor asset pricing model," *Journal of Financial Economics*, Vol. 116, No. 1, pp. 1–22.
- Farboodi, Maryam and Laura Veldkamp (2020) "Long-Run Growth of Financial Data Technology," *American Economic Review*, Vol. 110, No. 8, p. 2485–2523, August.
- (2023) "A Model of the Data Economy." NBER Working Paper No. w28427.
- Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran (2021) "Where Has All the Data Gone?," *The Review of Financial Studies*, Vol. 35, No. 7, pp. 3101–3138.
- Farboodi, Maryam, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran (2024) "Valuing Financial Data," *The Review of Financial Studies*.
- Flesch, Rudolf (1948) "A new readability yardstick," *Journal of Applied Psychology*, Vol. 32, No. 3, pp. 221–233.
- Franzoni, Francesco and Martin C. Schmalz (2017) "Fund Flows and Market States," *The Review of Financial Studies*, Vol. 30, No. 8, pp. 2621–2673, 03.
- Gardner, John (2021) "Two-Stage Differences in Differences," *Working paper*.
- Gardner, John, Neil Thakral, Linh T. To, and Yap Luther (2024) "Two-Stage Differences in Differences," *Working paper*.
- Genaro, Alan, Jose Maria Liberti, Pedro A. C. Saffi, and Jason Sturgess (2025) "Product Complexity, Investor Experience, and Returns," *Working paper*.
- Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer (2013) "Human Capital and Regional Development," *The Quarterly Journal of Economics*, Vol. 128, No. 1, pp. 105–164.

- Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2015) "Money Doctors," *The Journal of Finance*, Vol. 70, No. 1, pp. 91–114.
- Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov (2006) "Predicting volatility: getting the most out of return data sampled at different frequencies," *Journal of Econometrics*, Vol. 131, No. 1, pp. 59–95.
- Giroud, Xavier and Holger M. Mueller (2010) "Does corporate governance matter in competitive industries?" *Journal of Financial Economics*, Vol. 95, No. 3, pp. 312–331.
- Goldfarb, Avi and Catherine Tucker (2019) "Digital Economics," *Journal of Economic Literature*, Vol. 57, No. 1, p. 3–43.
- Goodman-Bacon, Andrew (2021) "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, Vol. 225, No. 2, pp. 254–277.
- Gormley, Todd A. and David A. Matsa (2011) "Growing Out of Trouble? Corporate Responses to Liability Risk," *The Review of Financial Studies*, Vol. 24, No. 8, pp. 2781–2821.
- Gourio, François and Leena Rudanko (2014) "Can Intangible Capital Explain Cyclical Movements in the Labor Wedge?" *American Economic Review*, Vol. 104, No. 5, p. 183–88.
- Grossman, Sanford J. and Joseph E. Stiglitz (1980) "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, Vol. 70, No. 3, pp. 393–408.
- Gârleanu, Nicolae and Lasse Heje Pedersen (2018) "Efficiently Inefficient Markets for Assets and Asset Management," *The Journal of Finance*, Vol. 73, No. 4, pp. 1663–1712.
- Gómez, Juan-Pedro, Melissa Porras Prado, and Rafael Zambrana (2024) "Capital Commitment and Performance: The Role of Mutual Fund Charges," *Journal of Financial and Quantitative Analysis*, Vol. 59, No. 2, p. 727–758.
- Harris, Lawrence E., Samuel M. Hartzmark, and David H. Solomon (2015) "Juicing the dividend yield: Mutual funds and the demand for dividends," *Journal of Financial Economics*, Vol. 116, No. 3, pp. 433–451.
- Hartzmark, Samuel M. and Abigail B. Sussman (2019) "Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows," *The Journal of Finance*, Vol. 74, No. 6, pp. 2789–2837.
- He, Bianca, Lauren Mostrom, and Amir Sufi (2024) "Investing in Customer Capital," No. 33171.
- Hoberg, Gerard and Ekaterina Neretina (2024) "Do Trade Associations Matter to Corporate Strategies?" *Working paper*.
- Hoberg, Gerard and Gordon Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, Vol. 124, No. 5, pp. 1423–1465.
- Hoberg, Gerard, Kumar Nitin, and Nagpurnanand Prabhala (2017) "Mutual Fund Competition, Managerial Skill, and Alpha Persistence," *The Review of Financial Studies*, Vol. 31, No. 5, pp. 1896–1929.

- Hortaçsu, Ali and Chad Syverson (2004) "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *The Quarterly Journal of Economics*, Vol. 119, No. 2, pp. 403–456.
- Ibert, Markus, Ron Kaniel, Stijn Van Nieuwerburgh, and Roine Vestman (2017) "Are Mutual Fund Managers Paid for Investment Skill?," *The Review of Financial Studies*, Vol. 31, No. 2, pp. 715–772.
- Jones, Charles I. and Christopher Tonetti (2020) "Nonrivalry and the Economics of Data," *American Economic Review*, Vol. 110, No. 9, pp. 2819–58, September.
- Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng (2008) "Unobserved Actions of Mutual Funds," *The Review of Financial Studies*, Vol. 21, No. 6, p. 2379–2416.
- Kostovetsky, Leonard and Alberto Manconi (2018) "How Much Labor Do You Need to Manage Capital?."
- Kostovetsky, Leonard and Jerold B. Warner (2020) "Measuring Innovation and Product Differentiation: Evidence from Mutual Funds," *Journal of Finance*, Vol. 75, No. 2, pp. 779–823.
- Lancaster, Kevin J. (1966) "A New Approach to Consumer Theory," *The Journal of Political Economy*, Vol. 74, No. 2, pp. 132–157.
- Lerner, Josh and Peter Tufano (2011) "The Consequences of Financial Innovation: A Counterfactual Research Agenda," *Annual Review of Financial Economics*, Vol. 3.
- Loseto, Marco and Federico Mainardi (2023) "Oligopolistic Competition, Fund Proliferation, and Asset Prices."
- Lou, Dong (2012) "A Flow-Based Explanation for Return Predictability," *The Review of Financial Studies*, Vol. 25, No. 12, pp. 3457–3489.
- MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023) "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, Vol. 232, No. 2, pp. 272–299.
- Martin, Ian W.R. and Stefan Nagel (2022) "Market efficiency in the age of big data," *Journal of Financial Economics*, Vol. 145, No. 1, pp. 154–177.
- Massa, Massimo (2003) "How do family strategies affect fund performance? When performance-maximization is not the only game in town," *Journal of Financial Economics*, Vol. 67, No. 2, pp. 249–304.
- Menzio, Guido (2023) "Optimal Product Design: Implications for Competition and Growth Under Declining Search Frictions," *Econometrica*, Vol. 91, No. 2, pp. 605–639.
- Mihet, Roxana (2022) "Financial Information Technology and the Inequality Gap," *Swiss Finance Institute Research Paper*, No. 21-04.
- Moretti, Enrico (2004) "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions," *American Economic Review*, Vol. 94, No. 3, p. 656–690.
- (2010) "Local Multipliers," *American Economic Review*, Vol. 100, No. 2, p. 373–77.

- Morlacco, Monica and David Zeke (2021) "Monetary policy, customer capital, and market power," *Journal of Monetary Economics*, Vol. 121, pp. 116–134.
- Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008) "Coarse Thinking and Persuasion*," *The Quarterly Journal of Economics*, Vol. 123, No. 2, pp. 577–619.
- Mullally, Kevin and Andrea Rossi (2025) "Moving the Goalposts? Mutual Fund Benchmark Changes and Relative Performance Manipulation," *The Review of Financial Studies*, Vol. 38, No. 4, p. 1067–1119.
- Nike (2025) "Nike Cookie Policy," Technical report.
- Obizhaeva, Olga (2024) "Does Search Engine Visibility Help ETFs Attract Flows?" *Working paper*.
- Pellegrino, Bruno (2024) "Product Differentiation and Oligopoly: a Network Approach," *American Economic Review (forthcoming)*.
- Previtero, Alessandro and Ran Xing (2025) "Beyond Performance: Mutual Funds, Non-Alpha Services, and the Value of Financial Advisors," *Working Paper*.
- Pástor, Luboš, Robert F. Stambaugh, and Lucian A. Taylor (2015) "Scale and skill in active management," *Journal of Financial Economics*, Vol. 116, No. 1, pp. 23–45.
- (2022) "Dissecting green returns," *Journal of Financial Economics*, Vol. 146, No. 2, pp. 403–424.
- Reuter, Jonathan and Eric Zitzewitz (2006) "Do Ads Influence Editors? Advertising and Bias in the Financial Media," *The Quarterly Journal of Economics*, Vol. 121, No. 1, pp. 197–227.
- Roberts, Michael R. and Toni M. Whited (2013) "Chapter 7 - Endogeneity in Empirical Corporate Finance1," Vol. 2 of *Handbook of the Economics of Finance*: Elsevier, pp. 493–572.
- Roldan-Blanco, Pau and Sonia Gilbukh (2021) "Firm dynamics and pricing under customer capital accumulation," *Journal of Monetary Economics*, Vol. 118, pp. 99–119.
- Rosenthal, Stuart S. and William C. Strange (2004) "Evidence on the Nature and Sources of Agglomeration Economies," *Handbook of Regional and Urban Economics*, Vol. 4, pp. 2119–2171.
- Rossi, Alberto G. and Stephen Utkus (2024) "The diversification and welfare effects of robo-advising," *Journal of Financial Economics*, Vol. 157.
- Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei (2020) "Marketing Mutual Funds," *The Review of Financial Studies*, Vol. 34, No. 6, pp. 3045–3094.
- Salop, Steven C. (1979) "Monopolistic Competition with Outside Goods," *The Bell Journal of Economics*, Vol. 10, No. 1, pp. 141–156.
- Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan L. Zhang (2025) "Generative AI and Asset Management," *Working paper*.
- Shive, Sophie and Hayong Yun (2013) "Are mutual funds sitting ducks?" *Journal of Financial Economics*, Vol. 107, No. 1, pp. 220–237.

- Sirri, Erik R. and Peter Tufano (1998) "Costly Search and Mutual Fund Flows," *The Journal of Finance*, Vol. 53, No. 5, pp. 1589–1622.
- Stock, James H. and Motohiro Yogo (2005) *Testing for Weak Instruments in Linear IV Regression*, p. 80–108: Cambridge University Press.
- Sun, Yang (2021) "Index Fund Entry and Financial Product Market Competition," *Management Science*, Vol. 67, No. 1, pp. 500–523.
- Thakor, Anjan V. (2020) "Fintech and banking: What do we know?" *Journal of Financial Intermediation*, Vol. 41.
- van Binsbergen, Jules H, Xiao Han, and Alejandro Lopez-Lira (2022) "Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases," *The Review of Financial Studies*, Vol. 36, No. 6, pp. 2361–2396.
- Veldkamp, Laura (2011) *Information Choice in Macroeconomics and Finance*: Princeton University Press.
- Vokata, Petra (2021) "Engineering lemons," *Journal of Financial Economics*, Vol. 142, No. 2, pp. 737–755.
- (2025) "Juicing the Coupon Yield: How Banks Extract Rents from Behavioral Biases," *Fisher College of Business Working Paper*, No. 2025-22.

2 Are New Technologies Replacing the Information Produced by Financial Markets?¹

“The debates over market efficiency, exciting as they are, would not be important if the stock market did not affect real economic activity” Morck, Shleifer, and Vishny (1990)

2.1 Introduction

The emergence of new information technologies is profoundly changing how information is gathered, processed, and used in financial markets. Technology is making new sources of information available to market participants. Big data, analytics, and advances in artificial intelligence allow them to process this information faster and more efficiently, providing new and timely predictions or insights about firms’ fundamentals. A rapidly growing literature studies how the adoption of these data technologies affect investors’ information acquisition decisions, investment choices, and the resulting informational efficiency of stock prices (e.g., Foucault et al., 2025). These are important questions because firms (and other decision makers) rely on stock prices as a source of information to guide their decisions.² Hence, if they make markets more efficient, data technologies could indirectly improve firms’ decisions and bolster real economic efficiency.

Yet, mirroring investors, firms are also increasingly adopting big data analytics and artificial intelligence to inform their decisions (Baley and Veldkamp, 2025). These tools enable managers to extract more and better information from internal data sources (e.g., data collected on customers, products, employees, or suppliers) and obtain novel insights about their business opportunities.³ Our contribution is to show

¹This chapter is co-authored with Laurent Frésard (Università della Svizzera italiana and Swiss Finance Institute).

²See Bond et al. (2012) and Goldstein (2023) for surveys of the literature.

³For instance, Amazon analyzes web-traffic data to “improve their products” Amazon (2025), and Nike analyze similar data to “understand personal preferences” Nike (2025).

that by improving firms' internal information, data technologies *reduce* their reliance on stock prices as an external source of information. Thus, firms' growing use of data technologies lessens the informational role of stock markets in affecting real economic activities. Using the words of Morck, Shleifer, and Vishny (1990), our results suggest that the advent of data technologies, perhaps paradoxically, makes the "debates over market efficiency" less important, at least when it comes to providing information to decision makers.

We reach this conclusion by tracking firms' reliance on their stock price for information after they adopt specific technologies that collect and analyze data generated by visitors on their website (e.g., Google Analytics, LiveRamp, or Hotjar). For instance, data analytics tools improve predictions of future demand or costs based on customers' activities (Brynjolfsson and McElheran, 2016, or Goldfarb and Tucker, 2019). To understand why data technologies can alter managers' reliance on their stock price and how we can infer this effect from the data, we use a standard investment model in which a manager chooses how much to invest in a growth opportunity. The optimal investment increases in her expectation of the payoff of this opportunity, which depends on her internal information and the firm's stock price, which contains relevant but noisy information about the payoff (e.g., due to liquidity or sentiment shocks). Because her internal information is imprecise and the stock price is informative about the opportunity, the manager optimally conditions her decision on both her internal information and her firm's stock price (Chen et al., 2007).

The firm's investment is thus sensitive to its stock price, and the strength of this sensitivity increases with the informativeness of the stock price, that is, the precision of the price signal. Building on the idea that data technologies reduce uncertainty about future outcomes (Baley and Veldkamp, 2025), we posit that their use provides the manager with more precise internal information about the growth opportunity. A direct implication is that, all else equal, the manager relies more on this information and less on the stock price when she uses technologies to process internal data. The firm's investment becomes *less* sensitive to its stock price, an hypothesis we label the "replacement" effect.

Testing this hypothesis is challenging because the econometrician does not observe the manager's private internal information. If this information overlaps with the information contained in the stock price—which is likely—an OLS regression of the firm's investment on its stock price yields an estimate of the manager's "true" reliance on the stock price that is biased upward (Foucault and Fresard, 2014). Hence, as it is well-known in the literature, the estimated sensitivity could be positive even if

the manager ignores the stock market completely. Furthermore, the model predicts that the OLS estimate of the investment-to-price sensitivity should *not* change when the manager uses data technologies. With more precise internal information, the manager's true reliance on her firm's stock price decreases (the replacement effect). Yet, the bias increases because her private information weighs more in her investment decision. Since both effects offset exactly, we cannot infer the replacement effect by regressing firms' investment on their stock price.

We can however infer it by using an instrumental variable for the firm's stock price that is orthogonal to the manager's private information. Similar to Dessaint et al. (2019) we use the noise in the firm's stock price generated by liquidity shocks as an instrument. This source of variation is exogenous to the firm's growth opportunity by design, and unless the manager can perfectly identify the noise in the price, it should be orthogonal to her internal information.⁴ The replacement effect implies that the firm's investment should become *less* sensitive to the noise component of its stock price when the manager uses data technologies to improve her internal information. This effect can be estimated using an "instrumented" difference-in-differences approach that estimates whether the investment of firms that adopt data technologies becomes less sensitive to their *instrumented* stock price compared to that of non-adopting firms. Notably, although the decision to adopt data technologies may arguably be endogenous (e.g., firms adopting when they have better growth opportunities), we show that it does not bias our estimation of the replacement effect as long as firms' adoption is unrelated to the noise in their stock prices (which is both plausible and supported by the data).

We implement this approach using a large panel of 8,815 U.S. firms over the period 1980–2023. We identify whether and when firms adopt data technologies using BuiltWith, a company that tracks the installation and removal of various technologies by analyzing webpage code. We focus on 822 specific technologies (or plugins) related to collection of customer data, tracking, and A/B testing and we concentrate on retail-oriented firms because they are the most likely to extract insights about their growth opportunities from data generated on their website.⁵ In our sample, adoption of these tools rose from about 20% in 2005 to 40% in 2010, and nearly 95% by 2023. This pattern mirrors the broad diffusion of analytics and tracking technologies over

⁴Dessaint et al. (2019) show that noise in stock prices influences investment through managerial learning, and that managers cannot perfectly filter out the noise in stock prices.

⁵These tools allow firms to forecast product demand and learn about consumer preferences. For instance, industry reports describe how Adobe Analytics helps online retailers forecast demand ahead of the holiday season (see [news.adobe.com/news/\[...\]online-rising-yoy](https://news.adobe.com/news/[...]online-rising-yoy)). Similarly, firms use A/B testing—tools akin to randomized control trials—to compare alternative product pages. By observing which version generates higher engagement or conversion, firms can infer consumer preferences.

the same period. We show that adopting firms are more likely to issue sales guidance, and issue more precise sales forecasts, consistent with data technologies improving internal information about future demand.

We measure non-fundamental variation in stock prices using quarterly reinvestment of cash dividends by mutual funds. We construct the hypothetical purchases of each firm's stock implied by mutual funds reinvesting the cash received from dividends. These purchases are hypothetical in the sense that they are derived assuming that mutual funds reinvest the cash payments from firms' dividends keeping the distribution of their holdings constant (which we confirm in the data). Like [Schmickler and Tremacoldi-Rossi \(2023\)](#) and [Chen \(2024\)](#), we find that these dividend-induced purchases are associated with positive pressures on firms' stock prices.

Consistent with the replacement effect, the sensitivity of investment to the non-fundamental component of their stock —the “investment-to-noise sensitivity”—decreases for firms after adopting data technologies compared to non-adopting firms. This sensitivity decreases by about one half: a one-standard deviation increase in firms' instrumented stock price is associated with a 7.18% increase in investment before adoption, but 3.43% after adoption.⁶ Tracing the investment-to-noise sensitivity in event-time around adoption reveals no difference between adopting and non-adopting firms prior to adoption, and a replacement effect that materializes two years after adoption.

The results are robust to stacking the sample to account for heterogeneous treatment effect ([Baker et al., 2022](#)), to account for common variation within industries, and to alternative sample definitions or measurement choices, such as using Total Q from [Peters and Taylor \(2017\)](#). Moreover, falsification tests confirm that the replacement effect is *not* present after firms adopt technologies that are *unrelated* to tracking customer data, which dispel concerns that the replacement effect captures broader changes in firms' web strategy. This effect is also not present for non-retail firms, which are less likely to obtain internal information about future demand from customers' web traffic. Notably, the replacement effect partially reverses for adopting firms that are subject to the California Consumer Privacy Act (CCPA) after 2019. This regulation limits the personal information that firms can collect and process from customers in California. By limiting the scope of data collection, this act lowers the effectiveness of data technologies to improve managers' internal information. The magnitude of the replacement effect drops by about 12.5% for adopting firms become subject to the

⁶A one-standard deviation increase in the instrumented stock price raises investment by 2.3 percentage points relative to the unconditional mean of 0.32 (7.18%), and by 1.1 percentage points after adoption (3.43%).

CCPA.

We consider various alternative explanations. First, instead of improving managerial information about their fundamentals, data technologies might help managers to better identify the noise in their stock price. This would also reduce their investment-to-noise sensitivity. We confirm that managers can partly detect the noise since they trade *against* it (i.e., selling shares following dividend-reinvestment price pressure). Yet, inconsistent with a learning about noise channel, their propensity to trade against the noise does *not* change after they adopt data technologies. Second, the use of data technologies may render stock prices less informative for managers, *without* affecting their internal information. For instance, investors may have less incentives to produce information about firms using data technologies. Thus, the lower investment-to-noise sensitivity may simply indicate a reduction in the informational content of stock prices for managers. Although proxies for stock price informativeness decrease following the adoption of data technologies, the replacement effect remains present, with similar magnitude, for firms that do not experience any change in the informativeness of their stock price.

We also investigate the role of well-known channels that could explain changes in the relation between firms' investment and their stock price. First, we examine whether the adoption of data technologies improves firms' financing conditions (e.g., by making their profits less volatile), which could make their investment less sensitive to their stock price (Baker et al., 2003). Using various proxies for the cost and availability of external financing, we find no support for such a financing channel. Second, we investigate if the adoption of data technologies increases firms' market power, since this could also create a disconnect between firms' investment and their stock price (Hayashi, 1982, or Gutiérrez and Philippon, 2017). We find however no significant change in firms' market share, price-cost margins, or markups after adopting data technologies.

At a more aggregate level, we estimate that industries with more intensive adoption of data technologies exhibit a decline in the sensitivity of investment to Q over the past decades. While descriptive, this pattern suggests that the replacement effect may help reconcile part of the growing aggregate disconnect between financial markets and corporate investment documented by Gutiérrez and Philippon (2017) and Alexander and Eberly (2018).

Related literature. Our results add to, and connect, two distinct strands of research. First, we contribute to the literature analyzing how stock prices affect real decisions through their informational content (see Goldstein, 2023 for a recent survey). Morck, Shleifer, and Vishny (1990) first noted that secondary stock markets matter for the real

economy only if they influence firms' decisions. When they do, efficient stock markets reveal valuable information to decision markets, and improves capital allocation in the economy, a concept named "revelatory price efficiency" (Bond et al., 2012). We show that improvements in the quality of managers' internal information obtained from data technologies lessens how strongly their investment respond to their stock price. Although not necessarily impairing real economic efficiency, this replacement effect implies that the adoption of data technologies lowers the informational importance of the stock market as guiding firms' investment, despite the overall increase of stock prices' informativeness documented by Bai et al. (2016) and Davila and Parlatore (2025). In his recent survey, Goldstein (2023) asks "*whether alternative sources of information can replace information from the market.*" Our analysis directly speaks to that question.

Relatedly, our results shed light on the nature of information that stock prices reveal to managers. Since data analytics and customer-tracking tools improves internal information pertaining to future demand and customers' preferences, the replacement effect suggests that it is information about these dimensions that managers typically glean from stock prices. These results complement that of Dessaint et al. (2026) who document that, in the context of acquisitions, stock markets reveals primarily information about discount rates.

Second, our paper also adds to the growing literature examining how data technologies affect firms, investors, and financial markets. Recent work finds that the adoption of data technologies affects firms' decisions (e.g., hiring or products), organization, product markets, and performance.⁷ A complementary stream of studies indicates that the use of data technologies by investors affect their forecasts, capital collection and allocation, and modifies the efficiency of stock markets (see Foucault et al., 2025 for a recent survey).⁸ By showing that data technologies lowers firms' reliance on their stock price as a source of information, our analysis reveals that data technologies modifies the interactions between firms and financial markets. In a related paper, Cao et al. (2023) study one facet of these interactions by showing that firms adjust their disclosure when investors use data technologies to read and analyze firms' regulatory filings. To the best of our knowledge, the replacement effect

⁷For instance, the use of data technologies improve firms' productivity in certain tasks (Brynjolfsson et al., 2023), increase sales growth (Babina et al., 2024b), investment efficiency (Ferracuti et al., 2024) and innovation (Cockburn et al., 2018). These technologies also alter firms' internal organizations (Labro et al., 2023, and Babina et al., 2023), risk (Babina et al., 2024a), disclosure practices (Blankespoor et al., 2022), and reshape product markets concentration (Lu et al., 2025).

⁸For instance, investors' use of data technologies affects their forecasting (Dessaint et al., 2024, and Chi et al., 2025), their capital collection (Zanotti, 2025) and capital allocation (Abis, 2022, Bonelli and Foucault, 2025, or Bonelli, 2025). The effects on stock market efficiency are studied by Martin and Nagel (2022), Dou et al. (2025), and Dugast and Foucault (2025).

is novel and suggests that the informational efficiency of financial markets, which is potentially bolstered by data technologies, may be becoming a second-order driver of real economic efficiency.

Our paper further contributes to a growing methodological effort to measure firms' adoption and usage of data technologies (see Babina (2026) for a review). Recent work uses survey data, the hiring of specialists in data-related areas, the production of patents related to data technologies, or mentions of these technologies in firms' communication. We complement these approaches by focusing on the actual adoption of data technologies from firms' website (see Zanotti, 2025 for a similar approach). While it captures specific technologies (e.g., data analytics) applied to specific datasets (gathered from customers' traffic activity), our approach enables us to measure firms' adoption precisely (e.g., its exact timing), and plausibly link this adoption to improvements in managers' internal information about future demand. This granularity is important to test the replacement effect.

2.2 The Replacement Hypothesis

To examine how the use of data technologies can replace the information produced by financial markets, we consider a simple model of investment in which a manager relies on internal data and the stock market as sources of information about an investment opportunity. Our objective is to (i) illustrate how the use of data technologies affects firms' reliance of stock market information to make decisions, (ii) underline the challenges to identify this effect empirically, and (iii) justify our identification strategy.

2.2.1 The Investment Model

As in Subrahmanyam and Titman (1999), we consider an investment model with two dates, 1 and 2. A firm has a growth opportunity with an uncertain payoff at date 2 equal to

$$G(I, \tilde{\theta}) = \tilde{\theta}I - \frac{I^2}{2} \quad (2.1)$$

where I is the investment in the opportunity and $\tilde{\theta}$ is the marginal productivity of this investment (the firm's fundamentals). For instance, $\tilde{\theta}$ could represent the future demand for the firm's products or its cost of production. The fundamental $\tilde{\theta}$ is normally distributed, with mean zero (for simplicity) and variance $\sigma_{\tilde{\theta}}^2$.

At date 1, the manager chooses the investment that maximizes the expected payoff

conditional on her information, Ω_1 , about $\tilde{\theta}$. The optimal investment I^* solves

$$\max_I \mathbb{E}[G(I, \tilde{\theta}) | \Omega_1] = \mathbb{E}(\tilde{\theta} | \Omega_1)I - \frac{I^2}{2}, \quad (2.2)$$

so that,

$$I^*(\Omega_1) = \mathbb{E}(\tilde{\theta} | \Omega_1). \quad (2.3)$$

Thus, the optimal investment is equal to the manager's expectation of the marginal return on her investment. To form this expectation, the manager has access to two distinct sources of information (signals) at date 1. Her set of signals about $\tilde{\theta}$ is $\Omega_1 = \{s, p\}$ and, conditional on $\tilde{\theta}$, these signals are independent. We describe each in turn.

First, the manager has internal data (e.g., data about customers, products, or employees) and processes it to obtain a noisy signal about the fundamentals $\tilde{\theta}$. We denote this signal by $s = \tilde{\theta} + \chi$ where χ is an error term, that is normally distributed with mean zero and variance σ_χ^2 . We assume that the manager has imperfect internal information ($\sigma_\chi^2 > 0$), because, otherwise, she would not rely on the stock market for information. We denote the precision of the manager's signal as $\tau_\chi = \frac{1}{\sigma_\chi^2}$. Building on the idea that the manager relies on data technologies to reduce uncertainty about future outcomes (Baley and Veldkamp, 2025), we posit (in reduced form) that the precision of s is higher if she uses data technologies to collect and process data to anticipate $\tilde{\theta}$. For instance, the use of data analytics could improve the manager's prediction of future demand based on customers' data, or her prediction of future costs based on procurement and suppliers' data.

Second, the manager can obtain external information from the firm's stock price, p . We assume that $p = \tilde{\theta} + u$ where u is the noise in the stock price that is normally distributed with mean zero and variance σ_u^2 . This noise captures in reduced-form variations in the firm's stock price p that are unrelated to the fundamental $\tilde{\theta}$ (e.g., due to liquidity shocks). Hence, the stock price is only partially revealing $\tilde{\theta}$. Similar to Dessaint et al. (2019), we do not model investors' equilibrium trading decisions and market clearing conditions, but we simply assume that p is a noisy signal about $\tilde{\theta}$, whose precision increases with the amount of investors' information about $\tilde{\theta}$. In other words τ_u increases with (unmodeled) investors' information. In the terminology of Bond et al. (2012), τ_u captures the extent of "forecasting price efficiency" (FPE), that is, how accurately the stock price predicts the future value of $\tilde{\theta}$.

The firm's optimal investment given the manager's information at date 1 is

$$I^*(\Omega_1) = E(\tilde{\theta}|\Omega_1) = \underbrace{\frac{\tau_\chi}{\tau_\theta + \tau_\chi + \tau_u}}_{:=\alpha} s + \underbrace{\frac{\tau_u}{\tau_\theta + \tau_\chi + \tau_u}}_{:=\beta} p. \quad (2.4)$$

Equation (2.4) describes how the manager's signals at date 1 affect her investment decision. For simplicity, we define the relative weights on each signal as $\alpha = \frac{\tau_\chi}{\tau_\theta + \tau_\chi + \tau_u}$ and $\beta = \frac{\tau_u}{\tau_\theta + \tau_\chi + \tau_u}$ so that the optimal investment is given by $I^*(\Omega_1) = \alpha s + \beta p$. As the manager's internal signal about the fundamental $\tilde{\theta}$ is imperfect ($\tau_\chi < \infty$), she can improve her estimate of $\tilde{\theta}$ by relying on her firm's stock price as a source of information, that is, $\beta > 0$. In other words, the firm's investment is *sensitive* to its stock price, and the strength of this sensitivity is captured by β . Following [Bond et al. \(2012\)](#), β captures “revelatory price efficiency” (RPE). All else equal, this sensitivity increases when the stock price is more informative about $\tilde{\theta}$ since β increases in τ_u (see, for instance, [Chen et al., 2007](#)).

Equation (2.4) formalizes the idea that the use of data technologies can *replace* the information produced by the stock market for the manager, i.e., reduces RPE. Intuitively, if having access to new data (e.g., richer customers' data) provide the manager with more precise internal information, it will increase her reliance on internal data for investment decisions ($\frac{\partial \alpha}{\partial \tau_\chi} > 0$) and reduce her need to rely on the stock price as a source of information. Crucially, this argument requires that data technologies inform the manager about fundamentals that (at least partially) overlap with what she learns from the stock price. If instead these technologies inform the manager about dimensions orthogonal to what prices reveal, adopting them would not affect her reliance on the stock price. Unless the use of these technologies is associated with an improvement in the precision of the firm's stock price signal (τ_u) or reduction of the unconditional uncertainty of the fundamentals ($1/\tau_\theta = \sigma_\theta^2$), an improvement in τ_χ induces the manager to rely *less* on her firm's stock price to invest, that is, $\frac{\partial \beta}{\partial \tau_\chi} < 0$. We obtain the following prediction.

Proposition 2.1. (*The Replacement Hypothesis*) *All else equal, when the manager uses data technologies, her firm's investment is more sensitive to her internal signal s (α is larger), and simultaneously less sensitive to her firm's stock price p (β is lower).*

2.2.2 The Econometrician's Problem

Testing Proposition 2.1 is difficult because the estimation of β is challenging. Unlike the manager in the model, the econometrician does not perfectly observe

the private signal s (e.g., [Erickson and Whited, 2000](#)), creating an omitted variable problem. To understand the econometrician's inference problem, consider what he realistically observes to infer β from the data: a sample of firms' investment I^* and their stock price p . Using this sample, he might estimate the linear Equation (2.4) such that:

$$I^* = c + bp + \xi \quad (2.5)$$

where c is a constant and the normally distributed error term ξ contains the unobserved private signal of the manager, s , and some error ε (i.e., $\xi = \alpha s + \varepsilon$). The coefficient b measures the sensitivity of investment to the stock price. Estimated by Ordinary Least Squares (OLS), the coefficient $\hat{b}^{OLS}$ is a biased estimate of true sensitivity β . Using Equation (2.4), it is given by

$$\hat{b}^{OLS} = \beta + \alpha \frac{Cov(s, p)}{Var(p)}, \quad (2.6)$$

where $\alpha > 0$ (because $\tau_\chi < \infty$). Since $Cov(s, p) = \tau_\theta^{-1}$ and $Var(p) = \tau_\theta^{-1} + \tau_u^{-1}$, we conclude that $\hat{b}^{OLS} > \beta$, that is, the OLS estimate is biased upward. The presence of this bias implies that we cannot test the replacement effect by tracking how $\hat{b}^{OLS}$ changes when the firm adopts data technologies (i.e., when τ_χ increases). Indeed, simple calculations yields $\frac{\partial \hat{b}^{OLS}}{\partial \tau_\chi} = 0$, that is, $\hat{b}^{OLS}$ is insensitive to changes in the precision of the manager's signal (for τ_θ and τ_u fixed). To see why, let's write $\hat{b}^{OLS}$ as

$$\hat{b}^{OLS}(\tau_\chi) = \underbrace{\frac{\tau_u}{\tau_\theta + \tau_\chi + \tau_u}}_{\beta(\tau_\chi)} + \underbrace{\frac{\tau_\chi}{\tau_\theta + \tau_\chi + \tau_u} \cdot \frac{\tau_\theta^{-1}}{\tau_\theta^{-1} + \tau_u^{-1}}}_{Bias(\tau_\chi)}. \quad (2.7)$$

When the manager has access to more precise internal data (τ_χ increases), the first term in Equation (2.7) decreases since the manager relies less on her firm's stock price when the precision of her internal signal improves—the true β decreases. Yet, the second term (i.e., the bias) increases with τ_χ because the managers' private signal gets more weight in her decision, which increases the bias. When τ_θ and τ_u are fixed, both effects offset exactly.

The model suggests however that the true β (and its changes with τ_χ) can be identified by instrumental variables. To understand the required properties of a prospective instrument, recall that the manager's private signal is $s = \tilde{\theta} + \chi$ and the price is $p = \tilde{\theta} + u$. Both sources of information are unbiased noisy versions of the fundamental. Thus, any variation in a firm's fundamentals might contemporaneously shift both the manager's internal signal and the price. This means that a valid

instrument cannot rely on fundamental variation. A prospective instrument z must be correlated with the firm's stock price p such that $Cov(p, z) \neq 0$ (i.e., satisfies the inclusion restriction), but uncorrelated with the error term ξ in Equation (2.5) such that $Cov(\xi, z) = Cov(\alpha \cdot s + \varepsilon, z) = 0$ (i.e., satisfies the exclusion restriction). Since the error term ξ contains the unobserved private signal s , z must be uncorrelated with the manager's private signal s , or more precisely, with the part of s that is *orthogonal* to the price p . Put differently, z should capture a portion of the firm's stock price p that does not overlap with the manager's signal s .

The model points to an instrument that captures variation in the firm's stock price p that is unrelated to the fundamentals $\tilde{\theta}$. By construction, variation in p that is related to $\tilde{\theta}$ will be correlated with the manager's private signal s , and therefore violates the exclusion restriction. However, an instrument that only correlates with the noise component of the stock price, u , satisfies the exclusion restriction. Imagine an instrument z that satisfies this condition such that $Cov(u, z) \neq 0$ and $Cov(\tilde{\theta}, z) = 0$. If the econometrician observes z , he can identify the true β by estimating Equation (2.5) by instrumental variables, such that

$$\hat{b}^{IV} = \frac{Cov(I^*, z)}{Cov(p, z)} = \frac{Cov(\alpha \cdot s + \beta \cdot p, z)}{Cov(p, z)} = \alpha \cdot \underbrace{\frac{Cov(s, z)}{Cov(p, z)}}_{=0 \text{ since } s \perp z} + \beta \cdot \frac{Cov(p, z)}{Cov(p, z)} = \beta. \quad (2.8)$$

Similar to Dessaint et al. (2019), the intuition behind z is as follows. If the manager relies on her stock price as a source of information ($\beta \neq 0$), her investment decision is sensitive to variation in p , whether this variation is fundamental ($\tilde{\theta}$) or non-fundamental (u). If the manager cannot perfectly distinguish these components of p , the replacement hypothesis (Proposition 2.1) predicts that her investment should become *less* sensitive to the component of p that is instrumented by z , i.e., the non-fundamental component. In contrast, if the manager could perfectly distinguish these components (i.e., perfectly filter out the noise in its stock price), her investment decision should be insensitive to the noise component of the stock price, whether she uses data technologies or not.⁹

2.2.3 Variation in τ_χ and Identification

To test Proposition 2.1, we exploit the adoption of data technologies as a proxy for an increase in firms' internal information (τ_χ) and estimate whether β decreases when the firm adopts these technologies compared to when it does not (yet). We

⁹See Dessaint et al. (2019) or Yan (2024) for evidence that corporate managers cannot perfectly distinguish fundamental from non-fundamental movements in stock prices.

can test the replacement hypothesis with an “instrumented” difference-in-differences specification such that:

$$I^* = c + b_1p + b_2\delta_{\tau_\chi} + b_3p \times \delta_{\tau_\chi} + \xi, \quad (2.9)$$

where δ_{τ_χ} is a binary variable that equals one if the firm adopts data technologies, and zero otherwise, such that $\tau_\chi|\delta_{\tau_\chi} = 1 > \tau_\chi|\delta_{\tau_\chi} = 0$. The replacement effect implies that the firm’s reliance on its stock price to decide on investment should be *lower* when it uses data technologies ($\beta_{\delta_{\tau_\chi}=1}$) than when it does not ($\beta_{\delta_{\tau_\chi}=0}$). Therefore, it predicts:

Prediction 2.1. (*Identifying the replacement effect*) When the manager uses data technologies, her firm’s investment becomes less sensitive to the non-fundamental component of her firm’s stock price p , that is, $\hat{b}_3^{IV} < 0$.

2.2.4 Endogenous adoption of Data Technologies

The adoption of data technologies is unlikely to be exogenous, and might depend on unobserved factors related to firms’ investment decisions (e.g., firms adopting when they have valuable growth opportunities). This potential endogeneity might in fact not be an issue. Because we need to instrument p with z to properly identify β , we estimate Equation (2.9) by instrumental variables, instrumenting p using z , and the interaction term $p \cdot \delta_{\tau_\chi}$ using the interacted instrument $z \times \delta_{\tau_\chi}$. Hence, as long as the decision to adopt data technologies is *not* correlated with z (e.g., firms adopt when their stock price is more noisy), the coefficient b_3^{IV} is an unbiased estimate of $\beta_{\delta_{\tau_\chi}=1} - \beta_{\delta_{\tau_\chi}=0}$.¹⁰ Therefore, in addition to the standard relevance and exclusion conditions in 2SLS, we have a third identification condition: a valid instrument should be uncorrelated with the decision to adopt a data technology. This condition is important as it adds on the required properties that a prospective instrument needs to satisfy to ensure proper identification.

2.3 Data and Methodology

¹⁰Precisely, $\hat{b}_1^{IV}$ is an unbiased estimate of $\beta_{\delta_{\tau_\chi}=0}$, and $\hat{b}_1^{IV} + \hat{b}_3^{IV}$ is an unbiased estimate of $\beta_{\delta_{\tau_\chi}=1}$. Therefore, $\hat{b}_3^{IV}$ is an unbiased estimate of the difference in β when the firm uses or not data-related tools.

2.3.1 Sample Construction

Our sample consists of U.S. public firms from CRSP-Compustat merged (annual data). We start from all U.S. firms in Compustat (fic=USA, loc=USA, and curcd=USA) with common stocks (shrcd 10 or 11) trading on AMEX, NYSE, or Nasdaq (exchcd equal to 1, 2, or 3) for the 1980-2023 period. We keep observations that have (i) non-missing information on total assets, sales, capital expenditures, property, plant and equipment (PP&E), and (ii) total assets and sales greater than \$1 million.

To test our replacement hypothesis, we need variation in managers' internal information. We use firms' adoption of technologies aimed to collect and analyze customers' data as a reasonable proxy for an increase in the precision of managers' internal data (i.e., digitized information). This section introduces the data. We later provide support for our interpretation of an increase in managers' internal information (see Section 2.3.2.3.4).

We use data from BuiltWith to measure firms' adoption of data-related technologies. BuiltWith is an alternative data provider that analyzes websites' source code and searches for specific patterns, such as HTML tags, that identify the presence of technologies. Every website includes different "technologies" (or equivalently "plugins"), which work as building blocks. For instance, a website may use "Google Maps" to display an interactive map (e.g., to show store locations) and/or "Adobe Analytics" to collect and analyze web visitors' data. All of these technologies leave identifiable traces in the source code of a website. Thus, inspecting the website's code allows one to observe which plugins are installed in a given point in time. BuiltWith continuously crawls websites to capture the installation (and removal) of various technologies starting January 2000. As of September 2024, BuiltWith covers 15,881 distinct technologies, divided into 34 distinct categories, such as Analytics (e.g., Google Analytics), Security (e.g., reCAPTCHA, which limit access to users suspected of being bots), or Payment (e.g., PayPal).

We focus specifically on the "Analytics" category. We carefully read the description of what these technologies actually do, and retain only 822 unique plugins designed to collect, process, and analyze customers data. These include, for instance, technologies labeled as "User Tracking and Analytics", "A/B Testing", "Social Media Tracking", or "Cloud and Customer Relationship". These plugins allow firms to gather rich information about customers—such as demographics, product interest, and browsing patterns—that managers can use to tailor product design and communication. For example, one of the technologies in our sample, Hotjar, is a data analytics plugin that records user interactions to help firms understand how users behave on their website, what they search, and where they spend more time.

Data Technology Name	Installation % (in 2023)	Description
Google Analytics	95.43	Users Tracking and Analytics
Google Analytics 4	59.12	Users Tracking and Analytics
Facebook Pixel	51.59	Social Media Tracking and Analytics
LinkedIn Insights	35.97	Social Media Tracking and Analytics
Bing Universal Event Tracking	27.17	Users Tracking and Analytics
Hotjar	24.91	Users Tracking and Analytics
Adobe Analytics	19.07	Users Tracking and Analytics
Twitter Analytics	18.72	Social Media Tracking and Analytics
Google Optimize 360	16.02	A/B Testing
RapLeaf	14.74	Users Tracking
Yahoo Web Analytics	14.69	Users Tracking and Analytics
Hubspot	10.76	Users Tracking and Analytics
Crazy Egg	9.73	Track and Visualize User Interaction
LiveRamp	9.58	Data Connectivity Platform
Yahoo Dot	8.50	Users Tracking and Analytics
Salesforce	8.40	Cloud and Customer Relationship

TABLE 2.1: **Main Data Technologies:** This table reports the main data technologies installed on firms' websites, as of December 2023. These plugins are aimed to collect and process website visitors' data. The second column shows the percentage of firms having the technology installed on its website with respect to the total number of funds, as of December 2023. The third column reports a short description of the technology's features.

In Table 2.1 we list the leading data technologies, as of December 2023, for our sample of U.S. retail firms. Google Analytics accounts for the lion's share of adoption, with around 95% of retail firms installing it.¹¹ Other leading plugins include RapLeaf, which allows for detailed user tracking, as well as LiveRamp, which integrates big data across platforms. All technologies in Table 2.1 are designed to extract signals from web traffic visits. These features are not limited to the most commonly used data analytics tools by firms in our sample. Appendix Figure 2.9 displays the word cloud extracted from the description of all these data technologies. Frequent terms include for instance "customer", "data", "analytics" and "tracking", consistent with their focus on collecting and analyzing customer data.

¹¹Appendix 2.7.4 shows that our results are not driven solely by Google Analytics (Table 2.19).

We merge BuiltWith data with CRSP-Compustat merged using firms' websites. In our main test, we concentrate on retail-oriented firms because they are the most likely to extract information about their potential customers (and growth opportunities) from data generated on their website. We identify these firms using the text-based approach of Frésard et al. (2019) linking product vocabularies from the Bureau of Economic Analysis (BEA) input-output tables to firms' 10-K product descriptions. The input-output tables record the purchases of "commodity" outputs (any good or service) by producing "industries" or "end-users" (including exports, personal consumption, or the government). Each commodity comprises a textual description (between 1 to 25 distinct words) that summarizes the nature of the good or service provided. Frésard et al. (2019) identify the specific words associated with the goods or services that are directly purchased by end-users, together with their economic intensity. Contrasting these words with those extracted from firms' 10-K business descriptions, they compute (for each firm-year) a score that measures the intensity with which each firm's products and service flows towards end-users. We define retail-oriented firms as those firms that have a score above 50% (at least once throughout the sample), and discard firms with a score below 50%.¹² From the 13,645 distinct firms in this set, 8,815 have a website that is ultimately covered by BuiltWith.

For each firm-year, we define the binary variable $DATA_{i,t}$ as equal to one if firm i has installed at least one data technology on its website in year t , and zero otherwise. This variable is the empirical equivalent to δ_{τ_χ} in Equation(2.9), and we use it to capture an increase in the precision of the manager's internal signal (τ_χ). Figure 2.1 plots the evolution of the average value of $DATA_{i,t}$ by year, tracking the adoption of data technologies. The adoption of data technologies began in 2005, with the adoption rate reaching 40% in 2010 and 80% in 2015. Overall, the adoption of data-related technologies by public U.S. retail firms resembles the typical logistic curve associated with adoption of new technologies (e.g., Comin and Hobijn, 2010). In 2023, 96% of retail-oriented firms have data technologies installed on their website. Our empirical strategy exploits the heterogeneous adoption across firms and time.

2.3.2 Non-fundamental Shocks to Stock Prices

We identify non-fundamental variation in stock prices (the empirical analog of the instrument z in the model) using the demand pressure generated by the reinvestment of cash dividends received by mutual funds from their portfolio firms.¹³ Schmickler

¹²We show later that we obtain similar results if we define retail-oriented firms based on SIC codes associated with "retail" industries.

¹³In robustness results, we show that our main findings do not stem from this particular source of non-fundamental price variation. We find similar results using mutual funds' abnormal outflows

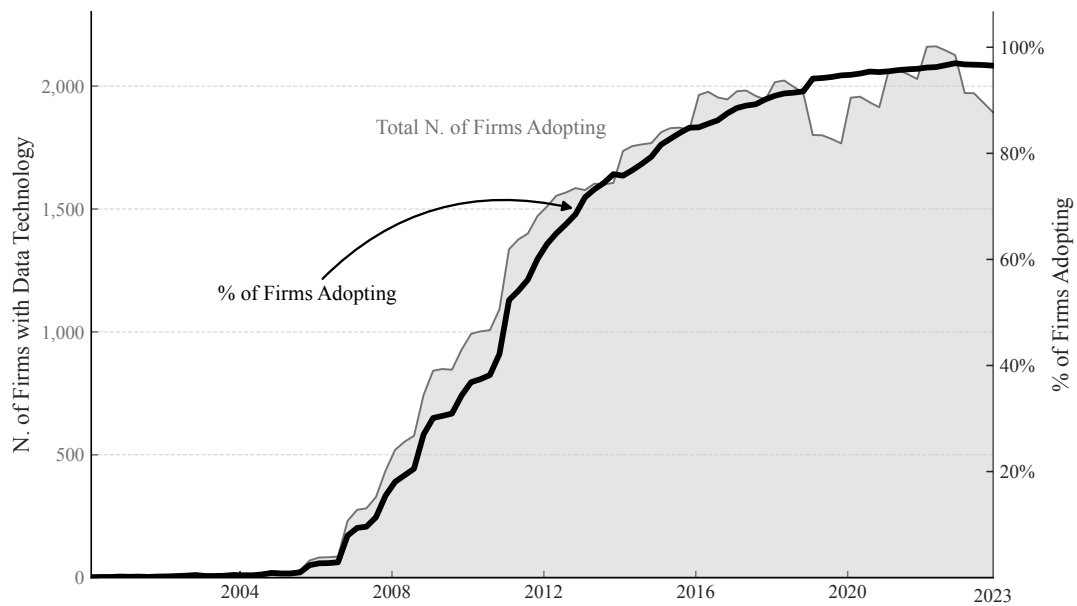


FIGURE 2.1: **Firms and Data Technology adoption.** This figure shows the adoption of data technologies in our sample. The data are from BuiltWith, which detects the installation and removal of various technologies by analyzing webpage code. The blue line (left axis) represents the number of funds with 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.

and Tremacoldi-Rossi (2023), Chen (2024) and van der Beck (2024) show that when funds reinvest cash dividends, they generate non-fundamental buying pressure for all stocks in their portfolio. The intuition behind this dividend-induced pressure can be illustrated with the following simple example. Consider a mutual fund holding an equally weighted portfolio consisting of three stocks, A, B, and C. At time t , stock A pays a cash dividend, while B and C do not. The fund receives cash from A and reinvests it proportionally across all holdings, increasing its positions in A, B, and C equally. Although A's dividend payment might reveal information about its fundamentals (e.g., its future performance), the increased demand for B and C is plausibly unrelated to their fundamentals. We rely on this demand pressure to capture variation in stock price that is plausibly unrelated to corporate managers' information about their firm's fundamentals ($\tilde{\theta}$).

Following Schmickler and Tremacoldi-Rossi (2023), we use quarterly holdings of mutual funds from Thomson Reuters and retain all U.S. funds between 1980 and 2023. We consider mutual funds because most have specific benchmarks and are thus more likely to reinvest the cash received from dividend payments in their existing holdings.

(inflows) as instrument for negative (positive) demand pressure, following the construction in Dessaint et al. (2026).

We obtain mutual funds' asset under management (AuM) from the CRSP Mutual Funds Database, and cash dividends paid by stocks from CRSP. Let $d_{i,q}$ denote the cash dividend per share paid by stock i in quarter q , and $S_{j,q^{ex}}(i)$ the number of shares of stock i held by mutual fund j , as of the ex-dividend date of the dividend paid in quarter q .¹⁴ For each mutual fund j , we construct the total cash dividend received in quarter q as

$$C_{j,q} = \sum_{i=1}^N d_{i,q} \cdot S_{j,q^{ex}}(i), \quad (2.10)$$

where N is the total number of stocks held by mutual fund j in quarter q^{ex} . $C_{j,q}$ represents the dollar amount of dividends received by mutual fund j in quarter q . Then, we construct the *hypothetical* reinvestment purchase made by mutual fund j in stock i as $C_{j,q} \cdot w_{j,q-1}(i)$, where $w_{j,q-1}(i)$ denotes the weight of stock i in mutual fund j 's portfolio in quarter $q-1$. This product represents the (dollar) demand pressure from fund j into stock i , if fund j reinvests the cash dividends proportionally. Finally, we define $HDIT_{i,q}$ as the total *hypothetical dividend-induced trades* made by all mutual funds holding stock i , normalized by total assets ($A_{i,t}$):

$$HDIT_{i,q} = \frac{\sum_{j=1}^J C_{j,q} \cdot w_{j,q}(i)}{A_{i,t}}, \quad (2.11)$$

where J is the total number of mutual funds. To avoid including fundamental information coming from a firm's own dividend, we follow [Schmickler and Tremacoldi-Rossi \(2023\)](#) and construct $HDIT$ for stock i excluding dividends from stock i and from all firms in the same two-digits SIC industry, since they may reveal correlated fundamental information. We obtain $HDIT_{i,t}$ for each firm-year by taking the sum of $HDIT_{i,q}$ across quarters.¹⁵

Mimicking the approach of [Edmans et al. \(2012\)](#), $HDIT$ aggregates the hypothetical number of shares mutual funds purchase in response to receiving cash dividends, as if mutual funds were reinvesting all of the cash inflow from dividends in proportion to their existing portfolio weights. We do not use *actual* trades because they likely reflect fund managers' private information about firms' fundamentals. Instead, we assume that in response to inflows from dividends, mutual funds use all cash inflow to expand their existing holdings proportionally.

Similar to [Schmickler and Tremacoldi-Rossi \(2023, Table 2\)](#), we confirm in Ap-

¹⁴Stock dividends are paid to whom is holding the stock at market opening of the ex-dividend date, while the actual payment will be made several days later. In our sample, the dividend payment is made 43 days after the ex-date, on average. We denote the dividend payment quarter as q , and the associated ex-date q^{ex} .

¹⁵We obtain similar results when we take the average of $HDIT_{i,q}$ across quarters.

pendix 2.7.2 that mutual funds do reinvest dividends proportionally to their existing holdings. Furthermore, we find no evidence that mutual funds save part of the cash inflow from dividends as cash holdings (Appendix Table 2.15, columns (3) and (4)). As further evidence, we show that mutual funds, on average, do not immediately pass through the cash dividends they receive from their portfolio holdings to their own investors. In the U.S., mutual funds face a flat 4% excise tax on the difference between their annual distribution proceedings and their actual payouts to investors. To avoid incurring the 4% tax, mutual funds need to distribute to their investors before the end of the calendar year in which they received the cash dividends, or at most in January of the following calendar year.¹⁶ This creates incentive for funds to not distribute dividends to investors until the end of the year. Appendix Figure 2.6 shows this pattern in our sample of U.S. equity mutual funds. Panel A reports the share of total fund distributions that occur in each calendar month. The dashed line indicates what one should expect if mutual funds were to distribute to their investors uniformly during the year (i.e., $1/12 \approx 8.3\%$). However, more than 70% of total distributions occur in December, consistent with funds retaining dividend proceedings during the year and distributing only at year-end. Panel B, in Appendix Figure 2.6 shows that this concentration does not originate from seasonality in dividend payments: the distribution of cash dividends paid to funds across calendar months is approximately uniform.

By construction, the variation of *HDIT* for a given firm (in a given year) is entirely driven by the dividends paid by unrelated firms and is thus unlikely to reflect changes in investors' views about the fundamentals of that firm. In addition, it does not depend on choices made by fund managers about which stocks to buy (or sell) to invest the cash received from dividends. As such, *HDIT* is a plausible proxy for non-fundamental shocks to stock prices. In support of this claim, Figure 2.2, displays the relationship between mutual fund hypothetical dividend-induced trades and stock returns around the cash dividend payment. Since almost all stocks in the sample are hit at least once per year by demand pressure from dividend reinvestment, variation in *HDIT* is primarily along the intensive margin. Thus, we estimate a distributed lag model of stock returns on *HDIT* with leads and lags up to ten trading days around the dividend payment date (event-date), with stock and date fixed effects and standard errors clustered within date (Schmidheiny and Sieglöcher, 2023). Panel A of Figure 2.2 plots the cumulative coefficients, scaled by the standard deviation of *HDIT* to represent the price response to a one-standard-deviation *HDIT* shock. Stocks subject to dividend-induced buying pressure experience a price increase of

¹⁶Distributions paid in January of calendar year $y + 1$ are treated as paid on December 31 of year y , for tax purposes.

approximately 0.04% over the days following dividend payment.

As further evidence supporting our identification, we run a falsification test around the *announcement* of dividends that will be paid to mutual funds (and eventually reinvested) in later dates.¹⁷ Panel B of Figure 2.2 repeats the analysis using dividend announcement dates, instead of payment dates. Since no cash changes hands at the announcement day, we should observe no demand pressure around this date. The cumulative coefficients are indeed indistinguishable from zero throughout the event window, confirming that the price impact in Panel A is unlikely to originate from mechanical effects or sorting (e.g., Wardlaw, 2020).

We further show that *HDIT* is unlikely to capture economy-wide or industry-specific characteristics. Figure 2.3 displays the average value of *HDIT* across firms for each year in our sample and across two-digits SIC industries. We observe no obvious clustering in any particular industry.

Finally, we verify that *HDIT* does not explain firms' adoption decisions to adopt data technologies. An important condition for the identification of the replacement effect is that our instrument (z in the model) is not correlated with the (endogenous) decision to adopt data technologies (Section 2.3.2.2.4). A priori, there is no obvious reason why the reinvestment of dividends paid by unrelated firms held in the same institutional portfolios should influence a firm's decision to adopt a new technology. We nevertheless formally test this condition by regressing the adoption indicator on contemporaneous and lagged values of *HDIT*, including firm and year fixed effects. Table 2.2 shows that variation in *HDIT* does not predict whether or when firms adopt data-related tools. This result supports the condition underlying our identification: non-fundamental shocks to stock prices generated by dividend reinvestments are unrelated to the changes in the precision of managers' internal signals, τ_χ . Therefore, *HDIT* isolates variation in the non-fundamental component of prices that appears orthogonal to managers' internal information.

2.3.3 Empirical Specification

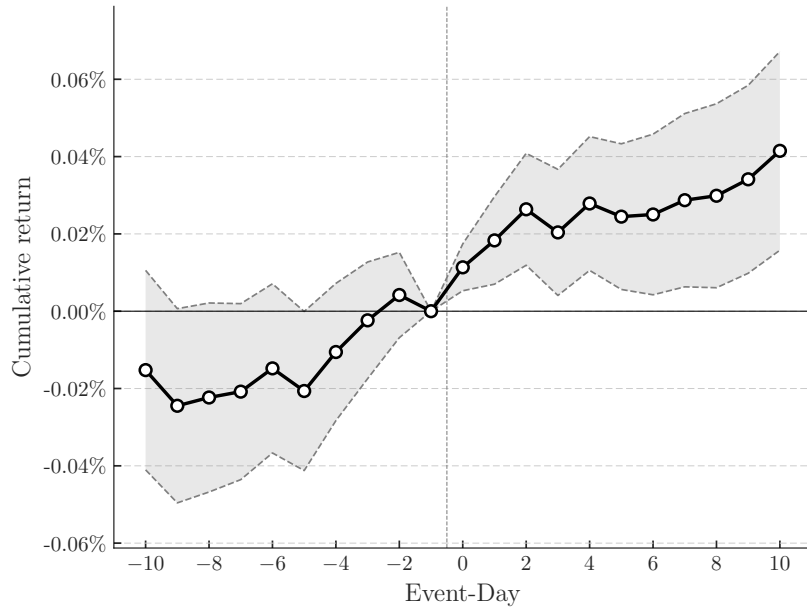
We specify the empirical counterpart of the model's difference-in-differences in Equation(2.9) as follows:

$$I_{i,t} = a_1 Q_{i,t-1} + a_2 \text{DATA}_{i,t-1} + a_3 Q_{i,t-1} \times \text{DATA}_{i,t-1} + \Gamma X_{i,t} + \gamma_i + \eta_t + \varepsilon_{i,t}, \quad (2.12)$$

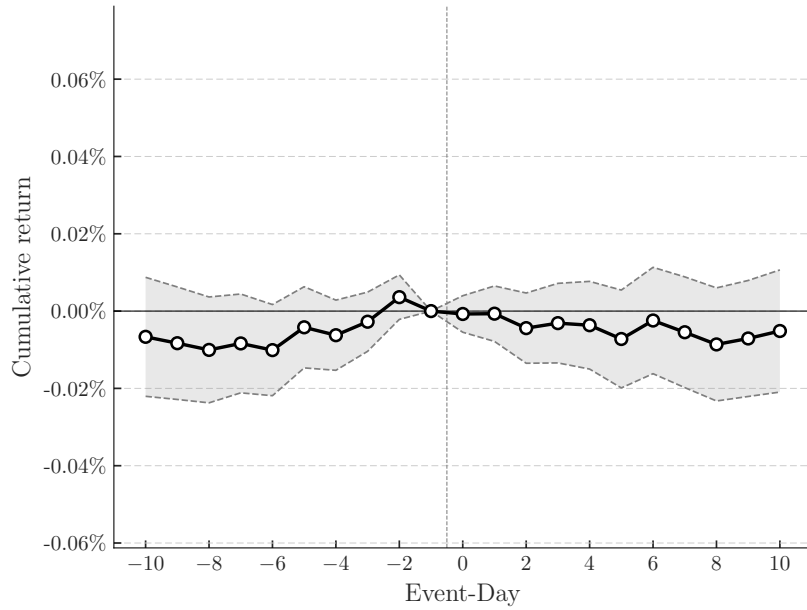
¹⁷The median (average) gap between a cash dividend announcement and payment dates in our sample is 31 (39) days. Appendix Figure 2.8 reports the distribution of gaps between dividend announcement and payment, in calendar days.

	$\mathbb{P}\{\text{Adoption}\}$					
	(1)	(2)	(3)	(4)	(5)	(6)
$\log(\text{Mkt Cap})_{i,t}$	0.004* (0.002)	0.002 (0.002)	0.001 (0.002)			
Inst. Ownership $_{i,t}$	0.035*** (0.013)	0.029** (0.014)	0.027** (0.012)			
Size $_{i,t}$	-0.001 (0.002)	0.000 (0.002)	0.002 (0.002)			
Age $_{i,t}$	0.001 (0.001)	0.000 (0.001)	0.000 (0.001)			
HDIT $_{i,t}$		0.004 (0.003)	0.003 (0.002)			
Peers' Adoption $_{i,t}$			0.919*** (0.024)			
$\log(\text{Mkt Cap})_{i,t-1}$				0.003 (0.002)	0.001 (0.002)	0.001 (0.002)
Inst. Ownership $_{i,t-1}$				0.029** (0.012)	0.026* (0.013)	0.027** (0.012)
Size $_{i,t-1}$				-0.001 (0.002)	0.000 (0.002)	0.002 (0.002)
Age $_{i,t-1}$				0.001 (0.001)	0.000 (0.001)	0.000 (0.001)
HDIT $_{i,t-1}$					0.003 (0.003)	0.001 (0.002)
Peers' Adoption $_{i,t-1}$						0.696*** (0.037)
Firm FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Obs.	127,518	110,972	110,972	115,780	101,030	101,030
Adj. R^2	0.818	0.817	0.840	0.815	0.813	0.826

TABLE 2.2: **Probability of adopting a Data Technology:** This table shows results of panel regression in which the dependent variable is the probability that a firm adopts a data-related tool. The regressors are the $\log$ market capitalization ($\log$ Mkt Cap), the share of total shares outstanding held by institutional investors (Inst. Ownership), a firm's size ($\log$ assets), age, Hypothetical Dividend Induced Trading (HDIT), and the share of peers with a data-related tool installed (Peers' Adoption within three-digits SIC industry). The annual sample is from 1980 to 2023. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.



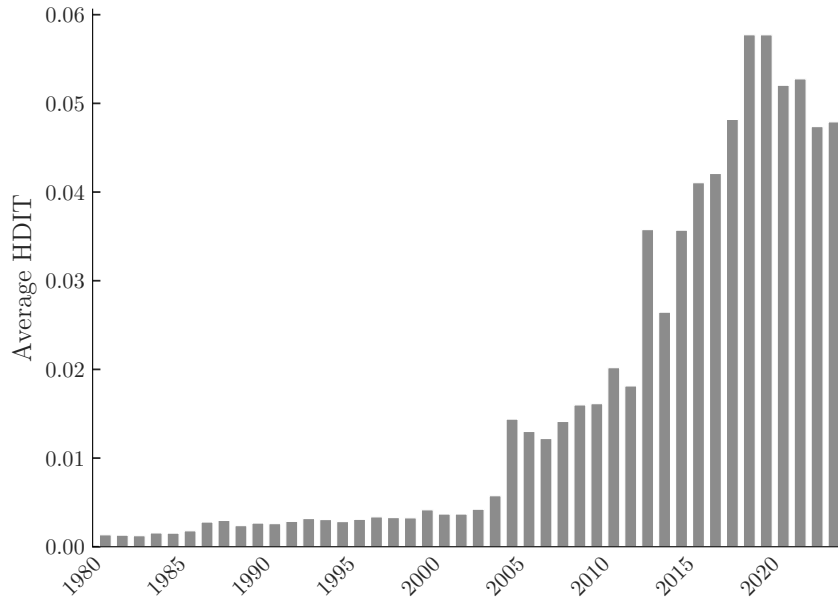
(A.) Dynamic Effect around Dividend Payments.



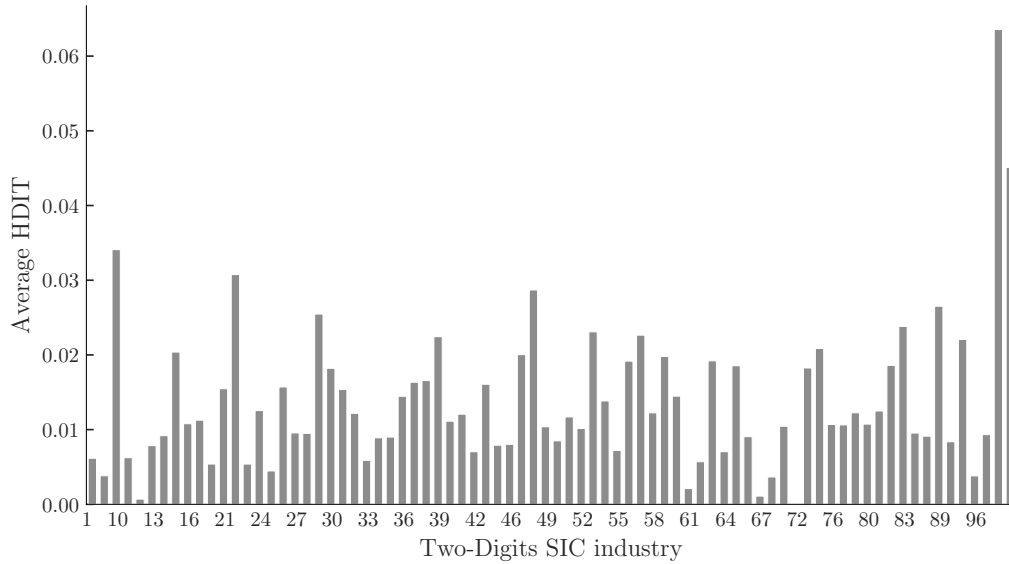
(B.) Dynamic Effect around Dividend Announcements.

FIGURE 2.2: Dynamics of the Effect of HDIT on Prices.. This figure plots the cumulative daily abnormal returns of stocks subject to Hypothetical Dividend Induced Trading (*HDIT*) price pressure. We estimate a distributed lag regression of daily stock returns on *HDIT* at leads and lags of up to ten trading days, with stock and date fixed effects and standard errors clustered within date and restricting the sample to stocks with *HDIT* larger than zero. Coefficients are scaled by one standard deviation of *HDIT* to represent the cumulative effect of one-standard deviation shock in *HDIT*. Panel A uses the actual dividend payment dates, while Panel B shows a falsification test using the dividend *announcement* dates. The shaded area denotes 95% confidence intervals.

where $I_{i,t}$ is the ratio of capital expenditure scaled by lagged fixed assets (property, plant, and equipment) in year t for firm i . $Q_{i,t-1}$ is the (normalized) stock price



(A.) Average HDIT by year.



(B.) Average HDIT by SIC2 Industry.

FIGURE 2.3: Hypothetical Dividend Induced Trading across Time and Industry. This figure plots the distribution of average Hypothetical Dividend Induced Trading (HDIT) by each year in our sample (Panel A), and by each SIC2 industry (Panel B).

defined as the market value of firm i (total assets minus book value of equity plus the market value of equity) divided by its total assets. The binary variable $DATA_{i,t-1}$ captures the adoption of data technologies. The vector $X_{i,t}$ includes standard control variables in investment models and variables capturing fundamental information about investment opportunities known to managers at the time they decide on investment. Specifically, we control for the one-year lagged values of the natural

logarithm of assets (firm “size”) and contemporaneous cash flows. In addition, we control for time-invariant firm heterogeneity by including firm fixed effects (γ_i), and aggregate fluctuations by including year fixed effects (η_t). Standard errors are clustered in two ways, by firm and by year. Appendix 2.7.1 describes the construction of all variables. We winsorize all ratios at 1% in each tail to reduce the effect of outliers. Table 2.3 presents the descriptive statistics.

	Obs.	Mean	Standard Deviation		P5	P25	P50	P75	P95
			sample	within-firm					
CAPEX	103,285	0.32	0.39	0.39	0.03	0.11	0.21	0.38	0.94
Size (<i>log Assets</i>)	103,285	5.90	2.15	0.89	2.68	4.27	5.75	7.37	9.74
Tobin’s Q	103,285	1.98	1.68	1.26	0.82	1.05	1.38	2.19	5.28
HDIT	103,285	0.02	0.03	0.02	0.00	0.00	0.00	0.02	0.08
Cash Flow	103,285	0.00	0.26	0.18	-0.49	0.01	0.06	0.11	0.20
<i>N. of Data Tech.</i>	103,285	1.39	3.84	-	0.00	0.00	0.00	1.00	9.00
DATA	103,285	0.25	0.44	-	0.00	0.00	0.00	1.00	1.00

TABLE 2.3: **Summary Statistics Annual:** 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 sample standard deviation (*sample*), the within-firm standard deviation (*within – firm*), the 5th (*p5*), 25th (*p25*), 50th (*p50*), 75th (*p75*), and the 95th (*p95*) percentiles. CAPEX is CAPX divided by PPENT, Cash Flows and Hypothetical Dividend Induced Trading (HDIT) are normalize by Total Assets. The variable *N. of Data Tech.* represents the total number of data technologies installed on the firm’s website in a given year, the variable DATA is a dummy equal to 1 if the firm-year observation has at least one data technology installed. All variables are winsorized at the 1% and 99% level. The annual sample is from 1980 to 2023.

We estimate Equation (2.12) by instrumental variables, using $HDIT_{i,t-1}$ to instrument $Q_{i,t-1}$ and the interaction between $HDIT_{i,t-1}$ and $DATA_{i,t-1}$ to instrument $Q_{i,t-1} \times DATA_{i,t-1}$. The coefficient of interest in Equation (2.12) is a_3 . It is the empirical equivalent of b_3 in Equation (2.9). It measures whether the sensitivity of firms’ investment to a non-fundamental component of their stock price (i.e., the *instrumented* component of Q) changes after they adopt data technologies. The replacement effect predicts that, as firms gather more precise internal information from data, this “investment-to-noise” sensitivity declines, that is, $a_3 < 0$. Although the adoption of data technologies starts in 2005, we estimate Equation (2.12) over the period 1980-2023. This sample choice follows Gardner et al. (2024), who recommend long pre-treatment periods in staggered adoption settings.¹⁸ Our sample begins in 1980 because it is the

¹⁸We show later that our conclusions are similar estimating Equation (2.12) on a sample that starts

first year in which we can construct our instrument for Q .

Similar to Dessaint et al. (2019) our test posits that managers do not perfectly observe the non-fundamental variation in their stock prices triggered by dividend-induced pressure. The econometrician can measure these changes ex post (possibly, a long time after they happened). However, our identification requires that these changes are not perfectly observed by managers when they happen, that is, managers cannot completely filter out the noise in their stock price. If they could, managers' beliefs (and investment decisions) would *not* be influenced by the effect of *HDIT* on stock prices (i.e., α_1 and α_3 in Equation (2.12) should be equal to zero). We show later (see Section 2.5.2.5.1) that, to some extent, managers detect part of the noise in their stock price. Yet, importantly, this detection ability is not affected by the adoption of data technologies.

2.3.4 Data Technologies and Managers' Information: Validation

The replacement effect rests on the hypothesis that the adoption of data technologies increases the precision of managers' internal signals about their fundamentals (τ_χ). Unfortunately, since we do not observe the precision of managers' signals, we cannot directly test this hypothesis. We provide however two pieces of evidence that support its validity, exploiting managers' guidance.

Managers issue guidance (i.e., forecasts) about their future sales and earnings on a voluntary basis, and are more likely to issue these forecasts when their internal signals differ from those of analysts or other market participants (Coller and Yohn, 1997). We posit that if data technologies improve the precision of managers' internal signal (τ_χ), they should issue guidance more frequently, and their forecasts should be more precise. We collect one-year ahead sales forecasts issued by managers from I/B/E/S Guidance from 2003 to 2023. We use one-year sales because managers most often release forecasts at that horizon. When forecasts are issued as a range instead of a point forecast, we use the upper bound of the range since it better reflects managerial expectations than the midpoint (Ciconte et al. (2014); Call et al. (2024)). We define a binary variable equal to one if a firm decides to release a sales forecast in a given year, and zero otherwise. We then measure the precision of each forecast computing the mean squared forecast error with the respect to the actual realization of sales, normalized by lagged sales.

Consistent with the idea that data technologies increase the precision of managers' internal information (τ_χ), Table 2.4 shows that managers are more likely to issue sales forecasts after they adopt these technologies. Columns (1) and (2) indicate in 1990.

	$\mathbb{P}\{\text{Issue Guidance}\}$		Guidance Squared Error	
	(1)	(2)	(3)	(4)
$\text{DATA}_{i,t}$	0.073*** (0.018)	0.088*** (0.023)	-0.039* (0.020)	-0.033*** (0.011)
Controls	×	✓	×	✓
Firm FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Outcome mean	0.111	0.111	0.074	0.074
Outcome SE	0.314	0.314	0.225	0.225
Obs.	47,611	40,302	9,808	9,638
Adjusted R^2	0.617	0.621	0.277	0.287

TABLE 2.4: **Manager Guidance and Data Technologies:** This table shows results of panel regression in which the dependent variable is the probability that a manager issue sales guidance (columns (1)-(2)), and the normalized squared error in the manager's guidance (columns (3)-(4)). The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. The control variables include a (lagged) firm's size ($\log$ assets) and contemporaneous cash flows over assets. Managers' guidance data are from I/B/E/S. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns (Gardner et al., 2024). The annual sample is from 2003 to 2023. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

that, after adoption, firms are about 8% more likely to release forecasts about their sales. Columns (3) and (4) further show that the adoption of data technologies is associated with a significant decline in managers' forecast errors. Taken together, results Table 2.4 support our hypothesis that data technologies are associated with improved managers' internal information about future demand.

2.4 The Replacement Effect

2.4.1 Baseline Results

To facilitate economic interpretation, in estimating Equation (2.12), we scale all independent variables by their within-firm sample standard deviation prior to estimation.¹⁹ This way, coefficients represent the change of investment in response to

¹⁹Because we include firm fixed effect in Equation (2.12), the source of variation used to estimate the coefficient is within-firm over time. For that reason, standardizing by the full sample standard deviation would inflate the economic magnitude of our results (Liu and Winegar, 2025). In this vein, in the 2SLS estimations, we standardize all covariates by their within-firm standard deviation. This choice of standardization explains why the standard errors in the 2SLS estimates (column (2)) are

a one-standard deviation change in each independent variable. Table 2.5 presents our main results. Estimated by 2SLS (i.e., instrumented with *HDIT*), the coefficient on *Q* is 0.023, with a *t*-statistic of 7.66. This estimate confirms that, overall, firms' investment is sensitive to the non-fundamental component of their stock price (e.g., Dessaint et al. (2019)). A one-standard deviation increase in this component is associated with a 2.3 percentage point increase in investment. Compared to the OLS estimation in column (3), this magnitude represents about a quarter of that of the non-instrumented coefficient on *Q* (0.089 with a *t*-statistic of 14.83), which captures both the fundamental and non-fundamental component of firms' stock price.

Consistent with the replacement effect (Prediction 2.1), the coefficient on $Q \times \text{DATA}$ estimated by 2SLS (i.e., α_3) is negative and statistically different from zero. The point estimate in column (2) is -0.012, with a *t*-statistic of 4. As predicted, the use of data technologies renders firms' investment *less* sensitive to the non-fundamental component of their stock price. After adopting these technologies, a one-standard deviation increase in the non-fundamental component of firms' stock price is associated with a 1.1 (i.e., 0.023-0.012) percentage point increase in investment, or about half the estimated magnitude prior to adoption (compared to non-adopting firms). The results are qualitatively similar (albeit smaller) in column (1) when we exclude the control variables.

Columns (4) and (5) presents the first-stage estimates. Confirming the results presented in Section 2.3.2.3.2, and in line with Schmickler and Tremacoldi-Rossi (2023), *HDIT* is positively and significantly related to *Q*. The magnitude of this effect is large, since a one-standard deviation increase in *HDIT* is associated with a 0.652 increase in *Q*. This increase corresponds to about half of the within-firm standard-deviation of *Q* in the sample. While the strength of the instrument *HDIT* is slightly lower after firms adopt data technologies (i.e., the coefficient on $HDIT \times \text{DATA}$ is negative), it remains substantial, alleviating weak instrument concerns.²⁰ This conclusion is further confirmed in column (5) as we observe a strong positive relation between $Q \times \text{DATA}$ and $HDIT \times \text{DATA}$, and considering the Kleibergen–Paap *F*-statistic of 36.71.

Notably, the coefficient on *DATA* itself is statistically significant, which suggests that firms adopting a data technology differ from non-adopters, after adoption. To ensure that the estimated change in investment-to-noise sensitivity is not merely reflecting these compositional differences, we verify in Section 2.5 that firms do not change

smaller than their OLS counterparts.

²⁰Note that this variation in the instrument strength in the first-stages has no effect on the magnitudes estimated in the second-stage because we standardize the instrumented *Q* and $Q \times \text{DATA}$ by their within-firm standard deviation.

Dependent variable:	CAPEX			Q	Q × DATA
	(1)	(2)	(3)	(4)	(5)
$Q_{i,t-1}$	0.017*** (0.002)	0.023*** (0.003)	0.089*** (0.006)		
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.005*** (0.001)	-0.012*** (0.003)	-0.034*** (0.009)		
$\text{DATA}_{i,t-1}$	0.116*** (0.023)	0.098*** (0.023)	-0.213*** (0.063)	0.069 (0.044)	1.182*** (0.049)
$\text{Size}_{i,t-1}$		-0.045*** (0.004)	-0.058*** (0.006)	-0.363*** (0.019)	-0.071*** (0.014)
$\text{Cash Flow}_{i,t}$		0.032*** (0.003)	0.034*** (0.003)	0.083*** (0.010)	0.014** (0.005)
$\text{HDIT}_{i,t-1}$				0.652*** (0.052)	0.024* (0.013)
$\text{HDIT}_{i,t-1} \times \text{DATA}_{i,t-1}$				-0.171*** (0.058)	0.360*** (0.024)
Estimator	2SLS	2SLS	OLS	First-stage	First-stage
Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.353	0.353	0.353	1.985	1.985
Outcome SE	0.460	0.460	0.460	1.772	1.772
Obs.	103,285	103,285	125,899	103,285	103,285
Adj. R^2	0.043	0.059	0.287	0.613	0.790
Kleibergen-Paap F-stat	36.53	36.71			

TABLE 2.5: **Main Results: Managers Learning and Data Technologies:** This table shows results of panel regression in which the dependent variable is the one-year-ahead firm's investment. The dependent variable in columns (1) to (3) is CAPEX. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

significantly along dimensions known to influence the sensitivity of investment to Q beyond the channel captured by our replacement hypothesis.

Figure 2.4 plots the evolution of the firms' investment-to-noise sensitivity prior and after the adoption of data technologies, compared to non-adopting firms. We

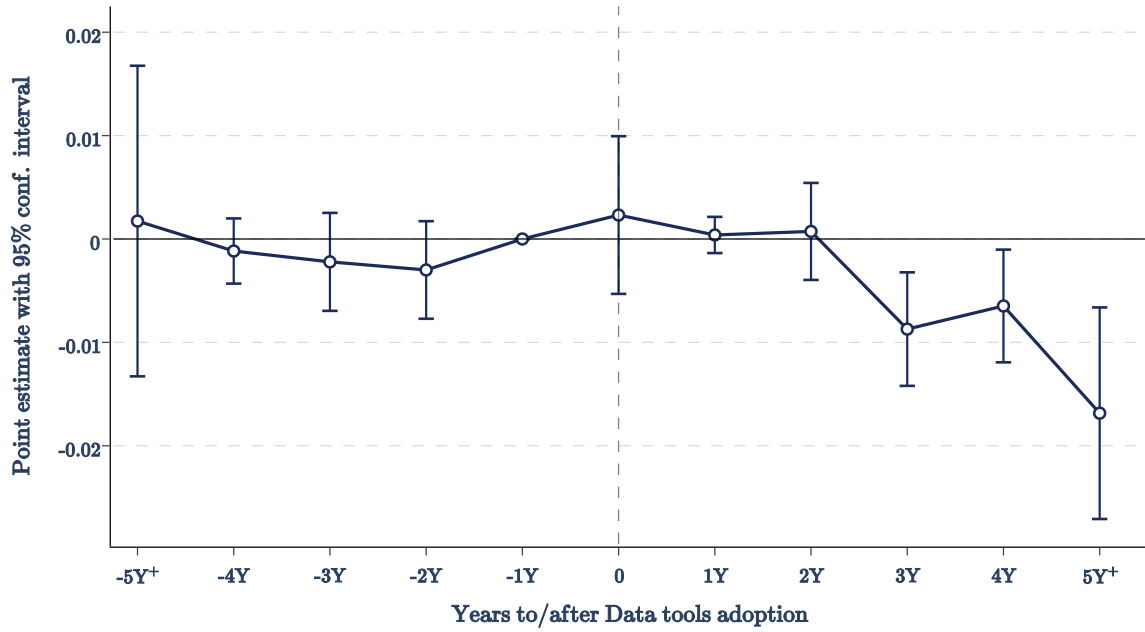


FIGURE 2.4: Dynamics of the Effect of Data on Managerial Learning. This figure shows results for the specification where Tobin's Q is instrumented by the Dividend Induced Trading (DIT) normalized by Total Assets. Each point represents the estimated coefficient on the interaction with each year before/after data technology adoption. The treatment is a dummy equal to one if a firm i has a data technology in place in year t ($DATA_{i,t}$). The control variables include a firm's size ($\log$ assets) in quarter t and contemporaneous cash flows over assets. The regression includes firm and year fixed effects, and the vertical bars represent the 95% confidence interval for the coefficient estimates. The year just before data technology adoption (-1) is the excluded category in the regression and is reported as zero in the figure. The rightmost estimates include all observations after 5 years from the adoption year. The annual sample is from 1980 to 2023.

set the year prior to adoption as the (excluded) reference period, and replace the binary variable $DATA$ in Equation (2.12) by event-time binary variables identifying the years surrounding firms' adoption (e.g., $DATA_{i,t}^{-2} = 1$ if firm i will adopt in two years, and 0 otherwise). In the years preceding adoption, we find no difference in the sensitivity of investment to the non-fundamental component of stock prices between adopting and non-adopting firms. Confirming that our results stem from the adoption of data technologies, the investment of adopting firms become less sensitive to non-fundamental changes in their stock price in the third year following adoption. While supporting the replacement effect, Figure 2.4 suggests that this effect is not immediate, perhaps because it takes time for firms to learn (or validate) valuable information from data collected from website traffic (i.e., improving the precision of the internal signal, τ_χ) after adopting the new tools.

We conduct further robustness checks. Given firms' staggered adoption of data technologies, we estimate Equation (2.12) using a stacked regression approach. Following Baker et al. (2022), for each year t , we create a "cohort- t " sample, retaining

only firms adopting data technologies in year t as well as firms that never adopt these technologies (in any year t) as control firms. We then estimate our main specification on the sample of stacked cohort samples, replacing firm fixed effects by firm $\times$ cohort fixed effects to account for the duplicity of control firms in the stacked sample. Table 2.6 indicates that changing the estimation approach does not modify our conclusion. We continue to observe a reduction in the sensitivity of firms' investment to noise after adoption of a data technology. Further, the economic magnitude of the estimated replacement effect matches that reported in Table 2.5. This is not by chance. In our setting, the total weight attached to "forbidden comparisons" (Goodman-Bacon, 2021) is relatively low because the main sample begins several years before the first adoption (treatment), which allows precise estimation of group effects (Gardner, 2021; Gardner et al., 2024). A similar logic applies to period (time) effects.

Dependent variable:	CAPEX			Q	Q × DATA
	(1)	(2)	(3)	(4)	(5)
$Q_{i,t-1}$	0.016*** (0.004)	0.021*** (0.004)	0.090*** (0.006)		
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.004** (0.001)	-0.008** (0.004)	-0.035*** (0.009)		
$\text{DATA}_{i,t-1}$	0.108*** (0.024)	0.103*** (0.024)	-0.077 (0.059)	0.171*** (0.062)	1.084*** (0.048)
$\text{Size}_{i,t-1}$		-0.040*** (0.006)	-0.043*** (0.006)	-0.395*** (0.021)	-0.024*** (0.006)
$\text{Cash Flow}_{i,t}$		0.026*** (0.004)	0.027*** (0.004)	0.067*** (0.012)	0.004** (0.002)
$\text{HDIT}_{i,t-1}$				0.718*** (0.059)	0.016*** (0.004)
$\text{HDIT}_{i,t-1} \times \text{DATA}_{i,t-1}$				-0.226*** (0.061)	0.345*** (0.022)
Estimator	2SLS	2SLS	OLS	First-stage	First-stage
Cohort × Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.363	0.363	0.363	1.976	1.976
Outcome SE	0.474	0.474	0.474	1.814	1.814
Obs.	309,468	309,468	354,853	309,468	309,468
Adj. R^2	0.044	0.056	0.253	0.580	0.813
Kleibergen-Paap F-stat	59.107	62.772			

TABLE 2.6: **Robustness of Main Results stacking:** This table shows robustness tests for the main results in the stacked sample. For each year in which at least one firm is treated, we maintain firm-year observations of untreated firms as of the event year (i.e., the control group include never treated and not-yet-treated firms) and stack all observations. The specification is equivalent to the one in the main results in Table 2.5. The dependent variable in columns (1) to (3) is CAPEX. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Tobin’s Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include cohort × firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Dependent variable:		CAPEX						
Robustness:	Industry×Year FE		Retail SIC Codes		Sample starts 1990		Interacted Controls	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Q_{i,t-1}$	0.021*** (0.003)	0.027*** (0.004)	0.015*** (0.003)	0.019*** (0.004)	0.017*** (0.002)	0.021*** (0.003)	0.020*** (0.003)	0.021*** (0.003)
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.007*** (0.001)	-0.015*** (0.004)	-0.006*** (0.002)	-0.012** (0.005)	-0.004*** (0.001)	-0.008*** (0.002)	-0.005*** (0.001)	-0.004*** (0.001)
$\text{DATA}_{i,t-1}$	0.094*** (0.019)	0.079*** (0.019)	0.107*** (0.033)	0.096*** (0.033)	0.113*** (0.023)	0.094*** (0.024)	-0.009 (0.063)	-0.061 (0.061)
$\text{Size}_{i,t-1}$		-0.043*** (0.004)		-0.033*** (0.005)		-0.041*** (0.005)	-0.049*** (0.006)	-0.048*** (0.005)
Cash Flow $_{i,t}$		0.030*** (0.003)		0.025*** (0.003)		0.028*** (0.003)	0.033*** (0.003)	0.031*** (0.003)
Interacted Controls	×	×	×	×	×	×	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	×	×	✓	✓	✓	✓	✓	×
Industry×Time FE	✓	✓	×	×	×	×	×	✓
Outcome mean	0.353	0.353	0.303	0.303	0.349	0.349	0.349	0.349
Outcome SE	0.460	0.460	0.393	0.393	0.454	0.454	0.454	0.454
Obs.	99,753	99,753	52,881	52,881	83,118	83,118	100,205	99,753
Adj. R^2	0.039	0.053	0.038	0.051	0.043	0.056	0.059	0.053
Kleibergen-Paap F-stat	36.622	36.987	40.930	41.066	37.382	37.657	37.698	37.780

TABLE 2.7: **Robustness of Main Results:** This table shows robustness tests for the main results in the paper. The dependent variable is CAPEX across all specifications. Columns (1) and (2) report results including Industry×Year fixed-effects, using two-digits SIC industry codes. Columns (3) and (4) show results using SIC codes to identify retail sector's firms in our sample (rather than "end-users"). Retail sector's three-digits SIC industry code are 080, 090, 200-211, 226-239, 251-254, 259, 265, 267, 283-284, 300-309, 321-323, 363-366, 393-396, 411, 481, 491-493, 503-504, 506, 508, 509, 514-599, 602-606, 611, 623, 641, 701-704, 720-726, 729, 751-753, 760-799, 801-809, 841-843, 849, 922. Columns (5) and (6) show robustness results to using 1990 as starting date for our sample. Columns (7) and (8) report results interacting controls. The table report results for the instrumental variable (IV) estimates, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). The regressors are a dummy equal to one if a firm i has a placebo technology in place in year t ($\text{PLACEBO}_{i,t}$). All regressors are standardized by within-firm variation. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

In addition, we show in Table 2.7 that our conclusions are robust to various modifications of the main specification. They continue to hold if we control for common variation within SIC 2-digits industries (with industry×year fixed effects), define retail sector firms based on their SIC codes as opposed to the text-based “end users” label, or start the sample in 1990. The results also remain unchanged when we interact all the control variables with DATA to capture potential indirect effects of adopting data technologies.

We further report in the Appendix that the main results in Table 2.5 are robust to alternative measurements of several key variables. We obtain similar results if we replace Q by the measure of Total Q that scales firm value by a measure of “total” capital that accounts for intangibles (Table 2.16), following Peters and Taylor (2017). Results are unchanged if we scale capital expenditures by total assets instead of fixed assets (Table 2.17), or if we consider investments in tangible and intangible assets (Table 2.18), using the total investment of Peters and Taylor (2017). Notably, we find no replacement effect for purely intangible investments. This type of investments (e.g., R&D) does not become less sensitive to noise after adopting data technologies. Our conclusions also continue to hold when we exclude Google Analytics (the most adopted technology) from the definition of DATA (Table 2.19).

We also verify that our results hold when using an alternative source of non-fundamental price variation. Following Dessaint et al. (2026), we instrument for non-fundamental stock price movements using mutual fund fire sales (purchases) driven by abnormal net outflows (inflows). Their instrument addresses known limitations in the standard constructions of flow-induced trading documented by Wardlaw (2020).²¹ We construct the instrument following the approach in Dessaint et al. (2026), and verify in Appendix Figure 2.10 that the first stage holds in our sample. Appendix Table 2.20 reports the results. The results are qualitatively consistent with the replacement hypothesis. Columns (1) and (3), yield coefficients on the interaction term between instrumented Q and DATA that are negative and statistically significant, supporting our main findings. However, when we include control variables in columns (2) and (4), the point estimates are negative but no longer significant. The loss of significance might reflect lack of statistical power or control variables partially absorbing the variation from the instrument in the second-stage.

²¹For this test, we restrict the sample start to 1993 as mutual fund net flows are reliably available in CRSP Mutual Funds only from March 1993 onward (Pástor et al., 2015).

2.4.2 Placebo Tests: Non-data Technologies and Non-retail Firms

Two falsification tests confirm that our main results stem from an improvement in the quality of managers' internal signals (τ_χ) upon adopting data technologies. First, we consider the adoption of technologies unrelated to the collection and analysis of customer data. From BuiltWith, we gather data on the adoption of 823 distinct tools related to digital agency (e.g., Wix Studio), registrar (e.g., technologies included in the hosting provider's registration), copyright (e.g., Copyscape), shipping (e.g., Shopify shipping). We restrict our attention to these four categories because they exhibit adoption rates that are similar to data-related tools, and the timing of their adoption has little overlap with that of data technologies (i.e., <10% overlap). This ensures that the results are (i) unlikely driven by a lack of power, and (ii) not capturing the simultaneous effect of data technologies.

We re-estimate Equation (2.12), replacing $DATA_{i,t}$ with the adoption of each of these "placebo" technologies. Table 2.8 reveals that adopting of these technologies does not modify firms' investment-to-noise sensitivity. The coefficients on the interaction between Q and adoption are insignificant in three specifications. It is negative and significant at the 10% confidence for registrar tools, but the estimated magnitude is small. Table 2.8 dispels the concerns that our results reflect an overall changes in firms' web strategy, or the adoption of any type of technology.

Second, we focus on non-retail firms (defined as those with a value of "end-users" below 50%), positing that they should benefit less from analyzing data from their website. Since their clients are not directly retail consumers, these firms should generate less information about future demand from the activities of potential customers on their website. Thus, web traffic data technologies should not be able to replace the information produced by financial markets for these firms. In line with this conjecture, Table 2.9 shows that the coefficient on the interaction between Q and $DATA$ is negative but not statistically significant. The investment-to-noise sensitivity of non-retail firms does not materially change after installing data technologies on their websites.

2.4.3 Regulatory Limitations of Data Access

The replacement effect stems from firms' ability to use new technologies for collecting and analyzing customers' data to obtain more precise signals about their growth opportunities (τ_χ). Thus, this effect should weaken when, after adopting data-related tools, firms face limitations in the amount and type of data they can collect from website traffic. To test this prediction, we exploit the limitations imposed

	CAPEX				
	Placebo:	Agency	Registrar	Copyright	Shipping
		(1)	(2)	(3)	(4)
$Q_{i,t-1}$		0.016*** (0.003)	0.016*** (0.003)	0.016*** (0.003)	0.016*** (0.003)
$Q_{i,t-1} \times \text{PLACEBO}_{i,t-1}$		-0.003 (0.020)	-0.003* (0.001)	-0.003 (0.029)	-0.003 (0.003)
Controls		✓	✓	✓	✓
Firm FE		✓	✓	✓	✓
Time FE		✓	✓	✓	✓
Outcome mean		0.354	0.354	0.354	0.354
Outcome SE		0.460	0.460	0.460	0.460
Obs.		103,285	103,285	103,285	103,285
Adj. R^2		0.050	0.050	0.050	0.050
Kleibergen-Paap F-stat		202.29	201.75	196.39	201.22

TABLE 2.8: **Placebo Technologies:** This table shows the results of panel regression in which the dependent variable is the one-year-ahead firm's investment. The dependent variable is CAPEX. The table reports results for the instrumental variable (IV) estimates, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). The regressors are a dummy equal to one if a firm i has a placebo technology in place in year t ($\text{PLACEBO}_{i,t}$). The placebo technologies are reported above each respective column: column (1) uses Digital Agency technologies as placebo (e.g., Wix Studio), column (2) uses Registrar-specific plugins (e.g., technologies included in the hosting provider's registration), column (3) uses Copyright plugins (e.g., Copyscape), and column (4) Shipping technologies (e.g., Shopify shipping). The control variables (omitted for brevity) include a lagged firm's size ($\log$ assets) and contemporaneous cash flows over assets. All regressors are standardized by within-firm variation. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

by the California Consumer Privacy Act (CCPA). This regulation was enacted in 2020 with the objective of protecting Californian customers' privacy. It requires that firms doing business in California disclose to customers what personal data they collect, how they use it, and gives them the option to prohibit the sale of their data to third-parties or request their deletion.²²

To assess whether the replacement effect weakens for firms subject to data limitations, we define $\text{CCPA}_{i,t}$, a binary variable that is equal to one if firm i is incorporated in California and year t is greater than 2019, and zero otherwise.²³ We

²²CCPA applies to any firm with business in California that meets at least one of three criteria: (i) annual (gross) revenues of \$25 million, (ii) collecting personal information of 50,000 customers or more, or (iii) having half of sales from selling customers' data.

²³The law was effective from June 2018, but firms had a "grace period" until July 2020.

Dependent variable:	CAPEX			Q	Q × DATA
	(1)	(2)	(3)	(4)	(5)
$Q_{i,t-1}$	0.016*** (0.004)	0.022*** (0.005)	0.026*** (0.006)		
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.003 (0.002)	-0.009 (0.009)	-0.014 (0.017)		
$\text{DATA}_{i,t-1}$	0.051 (0.031)	0.034 (0.032)	0.020 (0.032)	0.099* (0.055)	1.395*** (0.060)
$\text{Size}_{i,t-1}$		-0.038*** (0.004)	-0.037*** (0.004)	-0.193*** (0.018)	-0.042*** (0.011)
$\text{Cash Flow}_{i,t}$		0.022*** (0.003)	0.020*** (0.003)	0.096*** (0.011)	0.024*** (0.007)
$\text{HDIT}_{i,t-1}$				0.461*** (0.046)	0.021 (0.013)
$\text{HDIT}_{i,t-1} \times \text{DATA}_{i,t-1}$				-0.112** (0.051)	0.326*** (0.023)
Estimator	2SLS	2SLS	OLS	First-stage	First-stage
Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.303	0.303	0.303	1.799	1.799
Outcome SE	0.372	0.372	0.372	1.530	1.530
Obs.	43,547	42,103	41,601	45,142	45,142
Adj. R^2	0.061	0.081	0.072	0.657	0.826
Kleibergen-Paap F-stat	28.702	28.499			

TABLE 2.9: **Placebo Tests with “end-users” below 50%:** This table shows placebo tests on the sample of firms with “end-users” below 50% (i.e., non-retail firms). The dependent variable in columns (1) to (3) is CAPEX. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Total Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data tool in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

then augment Equation (2.12), interacting all variables with CCPA. In the spirit of a partial treatment “reversal”, the coefficient of interest is that on the triple interaction $Q \times \text{DATA} \times \text{CCPA}$. Estimated by 2SLS (i.e., instrumenting Q with HDIT), it captures how the sensitivity of investment to the non-fundamental component of stock price changes when firms’ using data-related tools face constraints in their data collection

and usage.²⁴ We predict a positive coefficient, consistent with firms relying more on their stock price for information when the quality of their internal information deteriorates due to data limitations.

The results in Table 2.10 support this prediction. Following the enactment of the CCPA, firms that use data-related tools and are subject to the CCPA experience a significant increase in their reliance on their stock price. In column (1), the coefficient on $Q \times \text{DATA} \times \text{CCPA}$ is positive and significant. In terms of magnitude, the estimates in Table 2.10 indicate that, after adopting data-related tools, the sensitivity of firms' investment to the non-fundamental component of their stock price decreases by about 1.6 percentage points, in line with our main results. However, the third row of Table 2.10 indicates that data-usage restrictions following the CCPA's enforcement in 2020 offset this effect by 0.2 percentage points for Californian retail firms using data-related tools. The size of this effect is economically reasonable and consistent with the magnitude of the replacement effect we document.

2.5 Alternative Channels and Implications.

In this section, we consider various alternative channels that could explain why firms' investment becomes less sensitive to the non-fundamental component of their stock price after they adopt a data technology. Overall, we find no consistent empirical support for these channels. We also discuss the implications of the replacement effects in the aggregate.

2.5.1 Learning about the Noise in Stock Prices

First, we consider the possibility that the reduced sensitivity of firms' investment to the non-fundamental component of their stock price arises not because of the replacement effect, but because the adoption of data-related tools improves managers' ability to detect the noise in their stock price. In that scenario, managers do *not* change their reliance on their stock price for information about fundamentals ($\frac{\partial \beta}{\partial \tau_\chi} = 0$), but because they can better distinguish the fundamental from the non-fundamental component of their stock price. In that case, their investment should be less sensitive to the noise in their stock price, that is, less "complying" with the instrument.

To assess this possibility, we examine the trading behavior of corporate insiders. To the extent that managers can identify the non-fundamental upward price pressure generated by hypothetical dividend-induced trades, they should opportunistically

²⁴In estimating this specification, we estimate four first-stages, for Q , $Q \times \text{DATA}$, $Q \times \text{CCP}$, and $Q \times \text{DATA} \times \text{CCPA}$.

Dependent variable:	CAPEX		
	(1)	(2)	(3)
$Q_{i,t-1}$	0.023*** (0.002)	0.025*** (0.002)	0.026*** (0.002)
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.020*** (0.004)	-0.016*** (0.003)	-0.019*** (0.003)
$Q_{i,t-1} \times \text{DATA}_{i,t-1} \times \text{CCPA}_{i,t-1}$	0.002*** (0.000)	0.002*** (0.000)	0.001*** (0.000)
$\text{DATA}_{i,t-1} \times \text{CCPA}_{i,t-1}$	0.061*** (0.010)	0.032*** (0.013)	0.029* (0.017)
$Q_{i,t-1} \times \text{CCPA}_{i,t-1}$	-0.047*** (0.003)	-0.046*** (0.004)	-0.034*** (0.006)
$\text{DATA}_{i,t-1}$	0.061*** (0.012)	0.054*** (0.012)	-0.106* (0.053)
Controls	×	✓	✓
Interacted Controls	×	×	✓
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
Outcome mean	0.354	0.354	0.354
Outcome SE	0.460	0.460	0.460
Obs.	97,383	97,383	97,383
Adj. R^2	0.045	0.061	0.061
Kleibergen-Paap F-stat	63.26	70.47	76.35

TABLE 2.10: **Managers Learning and Data Technologies after CCPA:** This table shows results of panel regression in which the dependent variable is the one-year-ahead firm's investment. All columns reports results for the instrumental variable (IV) estimates, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). The dependent variable is CAPEX. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), a dummy equal to one if firm i is subject to the California Consumer Privacy Act (CCPA) in year t ($\text{CCPA}_{i,t}$), and controls for firm-year characteristics (omitted for brevity). The CCPA became enforceable on January 1, 2020. The control variables include a (lagged) firm's size ($\log$ assets) and contemporaneous cash flows over assets. Column (3) includes control variables interacted with DATA and CCPA in all possible combination. All regressors are standardized by within-firm variation. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by state and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

and profitably trade against it (e.g., [Khan et al., 2012](#)). We collect insider trades from the Thomson Reuters insider filings database, and focus on all officers with decision-making authority over the operations of the firm, all board members and

beneficial owners with more than 10% of a company's stock. Following [Cohen et al. \(2012\)](#) we consider opportunistic (or non-routine) trades and measures opportunistic insiders' net sales in a given year either as insiders' annual (dollar) sales minus purchases, divided by their firm market capitalization (at the end of the year) or as the net number of shares sold.²⁵

²⁵Similar to [Cohen et al. \(2012\)](#) we require an insider to make at least one trade in each of the three preceding years to define him or her as either an opportunistic or a routine trader. We define a routine trader as an insider who placed a trade in the same calendar month for at least three consecutive years. Opportunistic traders are defined as everyone else (i.e., non-routine). We set insiders net sales to zero when firm do not have opportunistic trades in a given year.

Dependent variable:	Corporate Insiders' Net Sales				Corporate Insiders' Returns			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Q_{i,t}$			0.034*** (0.010)				-0.029** (0.012)	
$Q_{i,t} \times \text{DATA}_{i,t}$			-0.016 (0.010)				0.005 (0.008)	
$\text{DATA}_{i,t}$	-0.031 (0.023)	-0.011 (0.011)	0.046 (0.042)	-0.002 (0.016)	0.040 (0.028)	0.033 (0.022)	-0.033 (0.054)	-0.014 (0.015)
$\text{HDIT}_{i,t}$				0.021*** (0.007)				-0.026*** (0.007)
$\text{HDIT}_{i,t} \times \text{DATA}_{i,t}$				-0.007 (0.010)				0.015 (0.010)
$\text{Size}_{i,t-1}$		0.016*** (0.004)	0.002 (0.007)	0.009 (0.006)		-0.009*** (0.003)	0.007 (0.009)	0.009 (0.006)
$\text{Cash Flow}_{i,t}$		0.029*** (0.003)	0.030*** (0.005)	0.029*** (0.005)		-0.026*** (0.003)	-0.028*** (0.006)	-0.028*** (0.006)
Estimator	Staggered DiD		2SLS	OLS	Staggered DiD		2SLS	OLS
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓
Outcome mean	-0.010	-0.010	-0.010	-0.010	0.073	0.073	0.073	0.073
Outcome SE	1.016	1.016	1.016	1.016	0.441	0.441	0.441	0.441
Obs.	96,391	96,391	96,391	96,391	34,701	34,701	34,701	34,701
Adj. R^2	0.001	0.000	0.002	0.098	0.003	0.002	0.016	0.144
Kleibergen-Paap F-stat			41.778				34.725	

TABLE 2.11: **Corporate Insiders' contemporaneous Net Sales and Returns on Insider Trades:** This table shows results of panel regression in which the dependent variable is the total (nonroutine) net sales of corporate insiders (columns (1) to (4)) and 12-month excess return of corporate insiders for firm i in year t . Nonroutine trades follows definition in [Cohen et al. \(2012\)](#), while insiders include officers with decision-making authority, all board members and beneficial owners of more than 10% of a company's stock. Control variables include a (lagged) firm's size ($\log$ assets) and contemporaneous cash flows over assets. Regressors are standardized by within-firm variation. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. Standard errors are two-way clustered by state and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Table 2.11 indicates that the adoption of data technologies does not affect the intensity of insiders' opportunistic trades. The coefficients on DATA are not significant in columns (1) through (8). Columns (3) and (4) further indicates that insiders' net selling activity is sensitive to the non-fundamental component of their stock price. Consistent with existing research examining insider trading around downward price pressure generated by large mutual funds outflows (e.g., Khan et al., 2012, or Ali et al., 2011), insiders also appear to detect part of the non-fundamental component on their stock price: they sell significantly more shares when their stock price increases due to the buying pressure generated by hypothetical dividend-induced trades. However, the sensitivity of insiders' net sales to the non-fundamental component of their stock price does *not* change after they adopt data-related tools (columns (3) and (7)). For this reason, the negative effect of data-related tools on firms' investment-to-noise sensitivity is unlikely to stem from managers' increased ability to detect the non-fundamental component of their stock price analyzing their customers' data.

2.5.2 Fundamental Uncertainty and the Precision of the Price Signal

Second, recall that in our model, firms' reliance on their stock price is given by $\beta = \frac{\tau_u}{\tau_\theta + \tau_\chi + \tau_u}$ (see Equation (2.4)). So far, we have attributed the estimated decrease in β to an improvement in the precision of managers' internal information (τ_χ) obtained from the collection of new data—the replacement effect. Yet, Equation (2.4) suggests that a reduction of β could arise even in the absence of a replacement effect, that is, if $\frac{\partial \beta}{\partial \tau_\chi} = 0$. Indeed, Equation (2.4) indicates that a reduction of β could be observed if the adoption of data-related tools coincides with a reduction of the unconditional uncertainty of firms' fundamentals ($1/\tau_\theta = \sigma_\theta^2$) and/or a decrease in the precision of firms' stock price as a signal for managers about their fundamentals (τ_u).

To assess this possibility, we study whether proxies for the the uncertainty of firms' fundamentals and the informativeness of their stock prices change around the adoption of data-related tools. As proxies for $1/\tau_\theta$, we consider the volatility of firms sales, cash flows, and earnings (scaled by assets), computed using three-years rolling windows. As a proxy for τ_u , we consider the jump-ratio measure developed by Weller (2018), which compares the amount of a stock's price movement caused by the announcement of earnings to its total price movement during the pre-earnings period. A higher jump-ratio indicates prices adjust abruptly when earnings are released (prices are less informative), while a lower ratio indicates information is incorporated into prices more smoothly and continuously (prices are more informative). We also proxy τ_u by measuring the ability of firms' stock price to predict their future earnings, following Bai et al. (2016) (Appendix Table 2.21).

Dependent variable:	<i>vol</i> (Sales)	<i>vol</i> (Cash Flow)	<i>vol</i> (Earnings)	Jump Ratio (Weller, 2018)	
	(1)	(2)	(3)	(4)	(5)
$DATA_{i,t}$	-0.001 (0.005)	0.001 (0.003)	0.010* (0.006)	0.042*** (0.011)	0.038** (0.006)
Controls	✓	✓	✓	×	✓
Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.145	0.070	0.071	0.383	0.383
Outcome SE	0.181	0.125	0.313	0.339	0.339
Obs.	91,527	101,181	97,436	82,534	80,378
Adjusted R^2	0.408	0.626	0.607	0.176	0.178

TABLE 2.12: **Firm Volatility and Price Informativeness:** This table shows results of panel regression in which the dependent variables is a measure of firm volatility using 3-years rolling window (columns (1) to (3)) or stock price informativeness (columns (4) and (5)). The measures of firm volatility are the volatility of sales over total assets (column (1)), the volatility of cash flow over total assets (column (2)), and the volatility of earnings over total assets (column (3)). Columns (4) and (5) use the stock price jump ratio around earning announcements (Weller, 2018). A higher jump ratio is associated with less informative stock prices. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($DATA_{i,t}$), and controls for firm-year characteristics. The control variables include a (lagged) firm's size ($\log$ assets), Tobin's Q , and contemporaneous cash flows over assets. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns Gardner et al. (2024). The annual sample is from 1980 to 2023. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Table 2.12 shows that, all else equal, the adoption of data-related tools is not associated with a lower variance of firms' fundamentals. The coefficients on DATA are insignificant for the sales and cash flow volatility, and marginally significant but positive for earnings volatility. Hence, whereas data-related tools improve managers' signals about their firms' fundamentals (see Table 2.4), they do not alter the unconditional uncertainty of those same fundamentals (σ_θ^2). This channel is thus unlikely to explain our main result. In contrast, the adoption of data technologies is associated with a significant increase in firms' jump-ratio, that is, a *deterioration* of their price informativeness. We find a similar deterioration in column (5) when we look at stock prices' ability to predict firms' next year earnings. Notably, the reduction of price informativeness occurs despite the increased incidence of guidance provided by adopting firms (see Table 2.4), perhaps reflecting a crowding out of private information (e.g., Verrecchia, 1982, Diamond, 1985, or Morris and Shin, 2002).

To evaluate whether the effect of data-related tools on firms' investment-to-noise sensitivity solely stems from the reduction of τ_u , we focus on firms that do not experience a reduction of τ_u after they adopt data-related tools. We identify those firms computing the contribution of each firm to the average effect of data-related tools on the jump-ratio.²⁶ We then estimate our main specification on the subsample of firms for which the treatment effect on price informativeness is positive or zero. Table 2.22 confirms that our results do not originate solely from a deterioration in price informativeness. Even among firms experiencing an *improvement* in price informativeness (columns (1) and (2)), column (4) shows that the coefficient on $Q \times \text{DATA}$ estimated by 2SLS remains negative and statistically different from zero, with a similar economic magnitude. In other words, consistent with the replacement effect, after adoption of data-related tools firms' investment are less sensitive to non-fundamental variation of their stock price.

2.5.3 Financing Channel

Third, previous work shows that the investment of financially constrained firms is more sensitive to the non-fundamental component of their stock price because their equity price determines how cheaply they can raise funds (e.g., Baker et al. (2003) or Hoberg and Maksimovic (2015)). An increase in stock price may lower the cost of funds, leading to more investment (the financing channel). Thus, the decrease in firms' investment-to-noise sensitivity following their adoption of a data technology

²⁶The contribution of each firm to the average treatment effect on jump-ratio are obtained using the influence function method Cook and Weisberg (1982). We estimate the average contribution of each firm (DFBETA) and isolate those firms above the median, i.e., for which the effect of adopting data technologies on the jump-ratio is more negative.

could reflect reductions of firms' financial constraints if adoption lowers the cost and availability of financing. It might if, for instance, the use of data-related tools reduces information asymmetries as managers provide more credible disclosures (see Table 2.4) or improves the predictability of earnings for firms' outsiders.

To assess whether our results could reflect such a financing channel, we test whether firms' cost of financing drops after they adopt data technologies. We use the measures of financing constraints developed by [Hoberg and Maksimovic \(2015\)](#) and extended by [Linn and Weagley \(2023\)](#), which are based on textual analysis of firms' 10-Ks. In particular, we use their scores of the intensity of debt-market and equity-market constraints, where a higher score indicates more binding constraints. These scores are available for all firms since 1997, including firms that are not seeking access to financing, and as such, are arguably free of any selection bias. In addition, we rely on proxies for managers' perceived cost of capital and hurdle rate developed by [Gormsen and Huber \(2025\)](#) collected from firms earnings call transcripts, available since 2002.

Panel A: Equity Constraints						
Financial Constraints:	Equity Constraints		Debt Constraints		Cost of Capital	Hurdle Rate
	(1)	(2)	(3)	(4)	(5)	(6)
$DATA_{i,t}$	0.002 (0.003)	0.056*** (0.022)	-0.002 (0.001)	-0.010 (0.011)	0.000 (0.000)	0.000 (0.000)
Measure from:	HM (2015)	LW (2023)	HM (2015)	LW (2023)	Gormsen and Huber (2025)	
Controls	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	-0.006	-0.018	-0.001	0.018	0.091	0.118
Outcome SE	0.093	0.669	0.056	0.603	0.014	0.017
Obs.	38,624	81,402	36,622	81,401	29,456	29,455
Adjusted R^2	0.629	0.705	0.453	0.569	0.899	0.837

TABLE 2.13: **Financial Constraints and Data Technologies:** This table shows results of panel regression in which the dependent variables is a measure of financial constraints. Columns (1) to (4) use text-based measures developed in [Hoberg and Maksimovic \(2015\)](#) and [Linn and Weagley \(2023\)](#). Columns (1) and (2) refer to equity financial constraints, while columns (3) and (4) to debt constraints. In columns (1) and (3) we use measures from [Hoberg and Maksimovic \(2015\)](#) (HM (2015)) and columns (2) and (4) from [Linn and Weagley \(2023\)](#) (LW (2023)). In columns (5) and (6) the dependent variables is a measure of managers' perceived cost of capital from [Gormsen and Huber \(2025\)](#) (column (5)) and managers' perceived hurdle rate (columns (6)). The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($DATA_{i,t}$), and controls for firm-year characteristics. The control variables include a (lagged) firm's size ($\log$ assets), Tobin's Q , and contemporaneous cash flows over assets. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Overall, we find no evidence supporting a financing channel. Table 2.13 indicates that the access to external finance does not improve after firms adopt data-related tools. While we find no change in the intensity of debt-market constraints, column (2) indicates that adoption is associated with *higher* equity constraints. The coefficient on DATA is positive and significant with the measure of Linn and Weagley (2023), in line with the deterioration of price informativeness reported in Table 2.12, or with an increase in asymmetric information (e.g., if managers' guidance contains only part of the information they learn from data). The last two columns show no significant changes in managers' perceived cost of capital and hurdle rate after adopting data technologies.

2.5.4 Market Power Channel

Fourth, we examine whether the reduction of firms' investment-to-noise sensitivity could result from increased market power obtained through the use of data-related tools (the market power channel). Recent work suggests that the use of data enables firms to gain market shares and power, and affects product market dynamics (e.g., Eeckhout and Veldkamp, 2022; Babina et al., 2024b). At the same time, existing research argues and shows that market power renders firms' investment less sensitive to their stock prices (e.g., Gutiérrez and Philippon (2017)). This reduced sensitivity could occur because of a lower need for firms with market power to be responsive to demand shocks (reflected in the stock prices) or investment opportunities (because capacity expansion depresses product prices), or because of lower cost of financing due to more stable and predictable cash flows.

Inconsistent with this channel, Table 2.14 shows that the adoption of data-related tools does not lead to an increase in market power in our sample. In columns (1) and (2), we find no effect on firms' market shares (computed at the three-digits SIC industry level) or price-cost margins (defined as sales minus cogs, divided by sales, following Nickell (1996)). In column (3), we follow De Loecker et al. (2020) and focus on firms' markups. If anything, contrary to the market power channel, we observe a significant decrease in proxy for retail firms' markups after firms adopt technologies aimed to collect and analyze web traffic data.

2.5.5 Implications for Long-run Trends in Investment

Taken together, our tests provide strong empirical support for the replacement effect: the adoption of data-related technologies improves the precision of firms' internal information and reduces their reliance on their stock price as an external

Dependent variable:	Market Share	Price Cost Margin	DeLoecker et al. (2020)
	(1)	(2)	(3)
$DATA_{i,t}$	-0.009 (0.006)	-0.003 (0.032)	-0.344*** (0.094)
Controls	✓	✓	✓
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
Outcome mean	0.087	0.253	1.658
Outcome SE	0.210	0.663	1.428
Obs.	103,285	103,285	103,285
Adjusted R^2	0.670	0.604	0.580

TABLE 2.14: **Market Power and Data Technologies:** This table shows results of panel regression in which the dependent variables is a measure associated with market power. Columns (1)-(2) use a firm-year market share, computed as the share of sales with respect to the total sales of the three-digits SIC industry. Columns (3)-(4) use the firm-year measure from [De Loecker et al. \(2020\)](#), while columns (5)-(6) use the firm-year price cost margin computed as $(sales - cogs)/sales$. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($DATA_{i,t}$), and controls for firm-year characteristics. The control variables include a (lagged) firm's size ($\log$ assets), Tobin's Q, and contemporaneous cash flows over assets. All columns show estimates using difference-in-differences estimator robust to staggered treatment concerns ([Gardner et al., 2024](#)). The annual sample is from 1980 to 2023. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

source of information. This set of results is unlikely to reflect alternative channels. Figure 2.1 indicates that, since the mid-2000s, an increasing number of firms have adopted these new tools. Thus, our findings imply that, over time, corporate investment should have become less sensitive to stock prices. In this section, we provide suggestive evidence that the replacement effect might partly explain the increasing disconnect between firms' investment and Q documented by [Gutiérrez and Philippon \(2017\)](#) and [Alexander and Eberly \(2018\)](#).

Specifically, we obtain the evolution of the investment-to-price sensitivity by estimating the following specification for each two-digits SIC industry j :

$$I_{i,t} = d_1^j Q_{i,t-1} + d_2^j Q_{i,t-1} \times (t - t_0) + \Gamma^j X_{i,t-1} + \gamma_i + \eta_t + \varepsilon_{i,t}, \quad (2.13)$$

where i and t represent firms (operating in industry j) and years, and t_0 is 1985, such that $t - t_0$ represents an incremental year count after that year (i.e., the period 1980-1985 are benchmark years). Similar to Equation (2.12), we include firm and year fixed effect, and control for firms' size and cash flows. The coefficients d_2^j capture the trend in the investment-to-price for each industry. To relate these trends to the

adoption of data technologies across industries, we calculate for each firm the fraction of years for which DATA= 1, and average these fraction by industry.

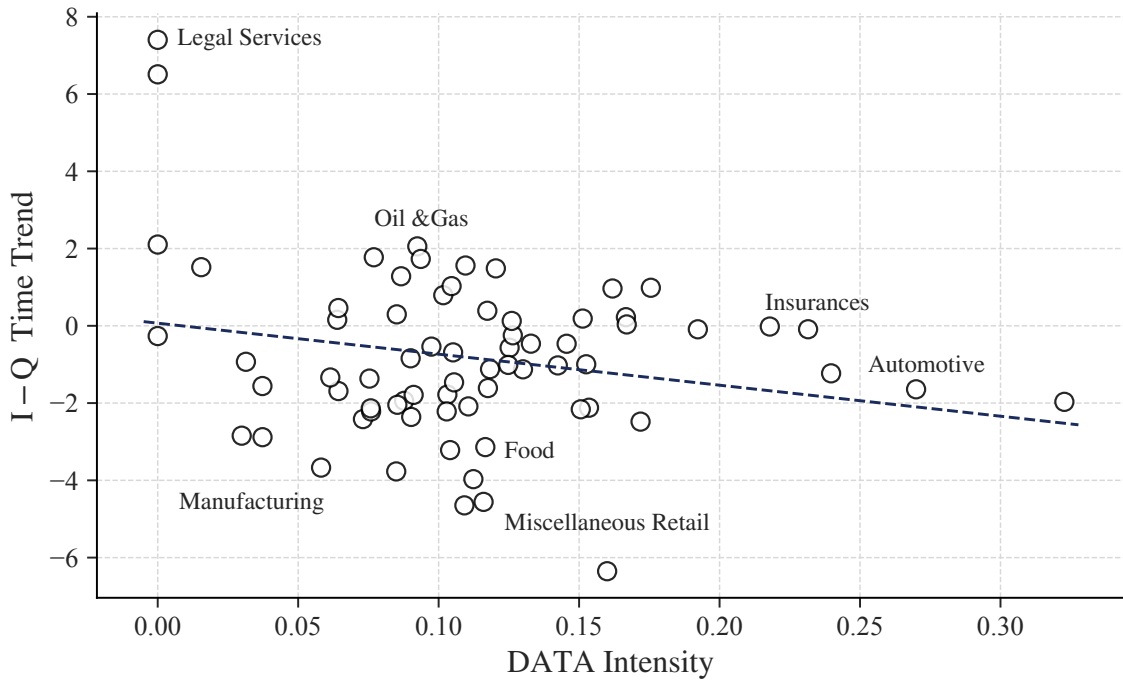


FIGURE 2.5: Secular Decline in I-Q Sensitivity and Data Technologies. This figure shows the scatter plot between the time-trend in I-Q sensitivity for each industry and the intensity of adoption within industry. For this plot we use the full sample of U.S. publicly listed firms (i.e., we do not restrict to retail-oriented firms.) In particular, we run the specification $I_{i,t} = \beta_1 Q_{i,t-1} + \beta_2 Q_{i,t-1} \times (t - t_0) + \Gamma X_{i,t-1} + \gamma_i + \eta_t + \varepsilon_{i,t}$, for each two-digit SIC industry, and plot on the y-axis the t-statistic for β_2 . The control variables include a firm's size ($\log$ assets) in quarter t and contemporaneous cash flows over assets. On the x-axis, we plot the average portion of data-adopters in our sample, within each two-digit SIC industry. The full annual sample is from 1980 to 2023.

Figure 2.5 plots the estimated trends (the t -statistics of d_2^j) in investment-to-price sensitivity for each industry against their intensity of data technology adoption. Confirming the growing disconnect between investment and Q documented by Gutiérrez and Philippon (2017) and Alexander and Eberly (2018), the estimates of d_2^j are negative for about 70% of industries (49 out of 71 industries). Notably, the estimated trends are negatively and significantly (at 10%) related to the intensity of data adoption across industries (coefficient of -0.803 with a t -statistic of 1.90). Overall, the results in Figure 2.5 are consistent with the prediction of the replacement effect that, over time, firms using new tools that increase their availability of internal data rely less on their stock price to decide on investment. Arguably, they could also be consistent with various alternative explanations. A more systematic analysis of the determinants of the aggregate decline in the sensitivity of firms' investment to their stock price is beyond the scope of this paper, but represents an interesting avenue for

future research.

2.6 Conclusions

We show that the diffusion of data technologies changes how firms learn from financial markets. When managers gain more precise internal signals from customer-generated data, they rely less on their stock price to guide real investment. Using non-fundamental variation in prices induced by mutual funds' dividend reinvestment, we document a sizable decline in investment-to-price sensitivity after firms adopt technologies that help them extract information about potential customers. The effect emerges only after adoption of these data technology, and is concentrated among retail-oriented firms—those for which customer data are likely most informative about future product demand. Consistent with data technologies replacing the informational role of financial markets, we find that the effect partially reverts when data-privacy restrictions, such as the California Consumer Privacy Act, constrain firms' ability to collect customer information.

Taken together, our results suggest that improvements in managerial information reduce the role of the stock market as a source of guidance for real decisions. This “replacement effect” does not imply worse investment; rather, it indicates that the source of information used in capital allocation is shifting from market signals towards firms' own datasets. At an aggregate level, this shift may help explain the gradual weakening of the link between investment and market valuations observed in recent decades. These findings open several further questions. For example, data technologies might similarly replace—or instead complement—market signals for other economic actors such as regulators, creditors, customers, or analysts. The answer to this question is not trivial, as managers' *internal* information set is very different from the one of other economic agents.

2.7 Appendix

2.7.1 Data Appendix

In this Appendix we describe the main dataset construction procedure. We use four main data sources: (i) Compustat Annual, (ii) CRSP Stock File Data, (iii) CRSP Mutual Fund Holdings, and (iv) BuiltWith for websites' technology installation/removal dates.

We construct the annual panel of U.S. public firms by merging Compustat Fundamentals Annual and CRSP through the standard CRSP–Compustat link (CCM). We filter the sample including only firms listed on the NYSE, NASDAQ, or AMEX (exchange codes 1–3) and classified as common shares (share codes 10 or 11). In our main text, we focus on retail-oriented firms because they are the most likely to extract signals about their growth opportunities from customer data from their website. We identify these firms using the text-based approach of Frésard et al. (2019) linking product vocabularies from the Bureau of Economic Analysis (BEA) input-output tables to firms' 10-K product descriptions. Frésard et al. (2019) identify the specific words associated with firms' goods or services that are directly purchased by end-users (together with their economic intensity). Then, they compute a score that measures the intensity with which each firm's products and service flows towards end-users. In our main sample, we keep only firms with a end-users score above 50% at least once throughout the sample (i.e., retail-oriented firms). Finally, we remove firms with negative or missing total assets and restrict the sample to observations after January 1980 to align with the availability of mutual fund holdings data.

To merge BuiltWith firms' website technologies with the main CRSP/Compustat panel, we rely on each firm's website URL from Compustat. We standardize URLs by removing prefixes and suffixes (e.g., "www.", "http://", "/index.html") and convert all domains to lowercase. For firms that change domains over time, we record the first and last year during which each website is active. Then, we merge each firm-year observation with the corresponding website technologies from BuiltWith. The final annual panel contains 103,285 firm-year observations, with 8,815 unique firms.

We construct *HDIT* using CRSP Mutual Funds holdings (s12) at the quarterly frequency. The procedure follows Schmickler and Tremacoldi-Rossi (2023). We use CRSP stock data at the daily frequency to align dividend ex- and payment dates with the corresponding fund-stock holdings.

We start from all U.S. mutual funds in the CRSP Mutual Funds database between 1980 and 2023. Using the s12 quarterly holdings, we retain equity positions (share

codes 10–11, exchange codes 1–3) and merge them with fund identifiers via *MF_LINK* tables provided by WRDS. We then compute portfolio weights each quarter using the fund’s total assets under management in equity.

Daily dividend information are from CRSP Distribution. We keep only cash dividends (distribution codes 1000–1399) and non-missing payment–ex-date pairs. We remove duplicate and require the payment date to be after the dividend announcement date (we drop 47 observations with this filter, about 0.01% of total stock dividend payment observations). For each dividend, we identify the trading day before the ex-date as the relevant date determining ownership. We obtain the dollar dividend amount by multiplying the dividend per share by shares outstanding on that date.

We assume holdings are constant within reporting quarter. For each stock-fund-day dividend event, we compute the dollar inflow to fund j from stock i in day d as:

$$\text{payFlow}_{i,j,d} = \text{divamt}_{i,d} \cdot q_{i,j,d}^{\text{ex}}, \quad (2.14)$$

where divamt is the dollar dividend paid by stock i in day d , and $q_{i,j,d}^{\text{ex}}$ is the fraction of outstanding shares of i held by fund j on the ex-dividend date. We then aggregate these “dividend-induced” inflows across all dividend-paying stocks in the fund’s portfolio to obtain the total dividend inflow received by the fund at each payment day. We then compute the dividend-induced inflow into each stock as

$$\text{inflow}_{i,j,d} = w_{i,j,d} \cdot \text{payFlow}_{i,j,d}, \quad (2.15)$$

with $w_{i,j,d}$ being the portfolio weight of stock i in fund j at that date (assuming constant holdings within the quarter).

We then aggregate across all funds holding stock i on each day to obtain the total hypothetical dividend-induced trading (*HDIT*). To remove any fundamental information, we exclude from $\text{HDIT}_{i,d}$ dividends paid by the stock itself and by firms in the same two-digit SIC industry. We then aggregate to the annual level using the average across the four quarters of year t , results are unchanged summing within year.

2.7.2 Dividend Reinvestment

The instrumental variable we use to identify plausibly non-fundamental demand shocks is the Hypothetical Dividend Induced Trading (*HDIT*) from mutual funds reinvestment of dividends. As detailed in Section 2.3.2.3.2, the underline assumption for this source of variation in stock prices is that mutual funds *reinvest* the cash they receive from dividends. We find empirical support for this assumption. First, we follow the approach used to validate similar assumptions in the literature (e.g., Lou (2012)). Specifically, we run the following pooled regression on our sample of mutual fund-stock-quarter holdings:

$$\frac{\Delta Shares_{j,q}(i)}{Shares_{j,q-1}(i)} = \alpha_j + \alpha_q(i) + \beta \frac{C_{j,q}}{AUM_{j,q}} + \varepsilon_{j,q}(i), \quad (2.16)$$

where $Shares_{j,q}(n)$ denotes the number of shares held by mutual fund j in quarter q , for stock i , $C_{j,q}$ is the total dividends payment received by fund j in quarter q (Equation 2.10), and $AUM_{j,q}$ is a fund's total assets under management. If mutual funds reinvest dividends proportionally to their existing holdings, we should find $\beta = 1$. If they do not reinvest any of the cash received from dividends, we should expect $\beta = 0$. We follow van der Beck (2024) in excluding large trades that exceed 100% of previous holdings ($\Delta Shares_{j,q}(i) > 1$) and excluding fund holdings below 0.1% of total assets under management. Columns (1) and (2) in Table 2.15 show the results in our sample. The estimates for β are close to 1, suggesting that mutual funds reinvest dividends proportionally in their existing assets. On average, when funds in our sample receive an inflow from dividends, they scale up their existing portfolio positions proportionally.

Second, we directly test whether mutual funds' cash holdings change after receiving dividends. If funds exert a certain degree of discretion in reinvesting dividend payments, they might use dividends to increase their cash buffer. To test this, we regress fund-quarter cash holdings on dividend repayments:

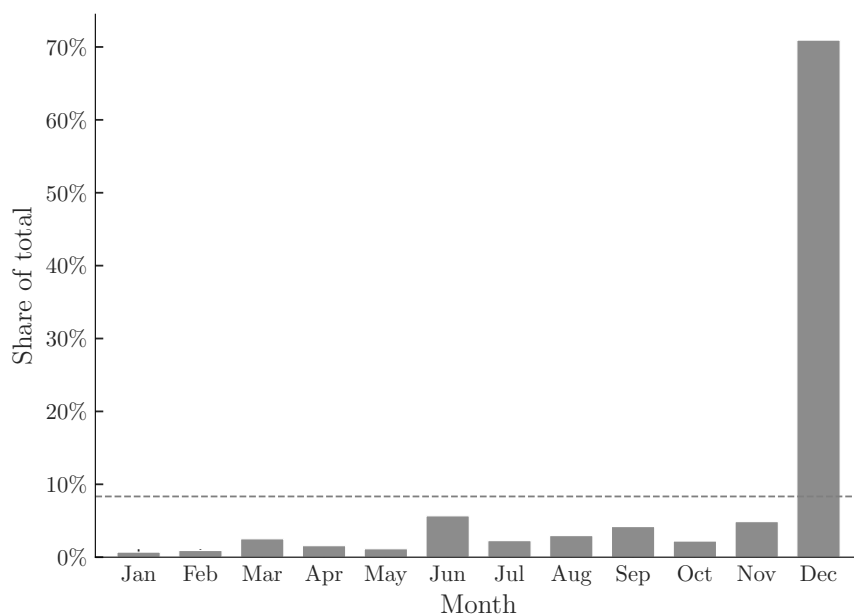
$$\frac{Cash_{j,q}}{AUM_{j,q}} = \alpha_j + \alpha_q + \delta \frac{C_{j,q}}{AUM_{j,q}} + \epsilon_{j,q}(i), \quad (2.17)$$

where $Cash_{j,q}$ denotes the cash holdings of fund j in quarter q . If mutual funds use dividend repayments to increase their cash buffer, we should find $\delta = 1$. On the other hand, if, on average, funds do not maintain dividend repayments as cash, we should expect $\delta = 0$. Columns (3) and (4) in Table 2.15 report the results. On average, fund managers do not increase their cash holdings when receiving dividend repayments.

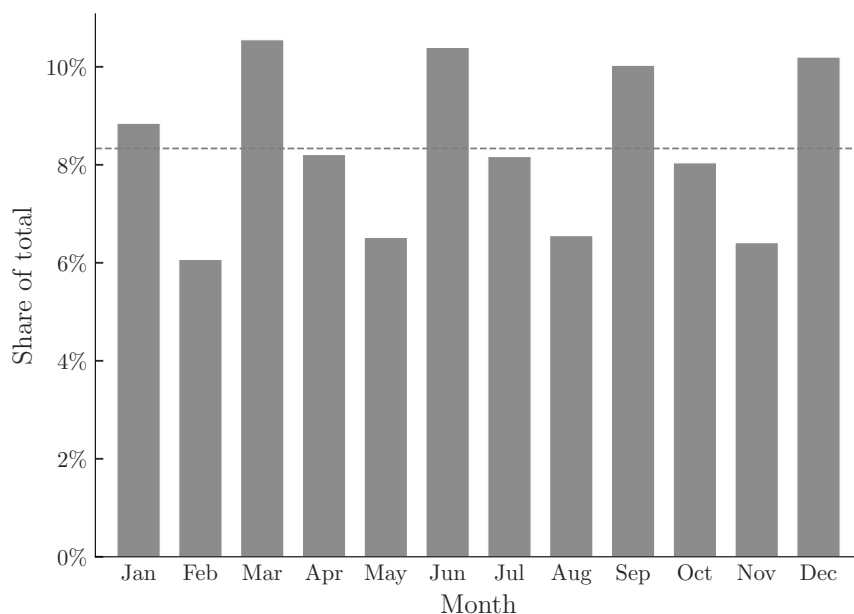
This evidence corroborates results in other papers using a similar approach to identify non-fundamental demand shocks.

	$\Delta Shares_{j,q}(i)/Shares_{j,q-1}(i)$		$Cash_{j,q}/AUM_{j,q}$	
	(1)	(2)	(3)	(4)
$C_{j,q}/AUM_{j,q}$	0.952*** (0.257)	0.952*** (0.257)	0.000 (0.000)	0.001 (0.000)
$Flow_{j,q}/AUM_{j,q}$		0.000 (0.000)		0.008*** (0.001)
Firm FE	✓	✓	✓	✓
Time FE	×	×	✓	✓
Stock×Time FE	✓	✓		
Outcome mean	0.219	0.219	0.034	0.034
Outcome SE	0.478	0.478	0.107	0.107
Obs.	23,441,516	23,441,516	1,867,721	1,867,721
Adj. R^2	0.515	0.515	0.573	0.573

TABLE 2.15: **Dividend Reinvestment:** This table shows empirical support for the assumption that mutual funds reinvest dividend payments. Column (1) reports results for the fund-stock-quarter regression in Equation 2.16. Column (2) further controls for mutual fund flows. Columns (3) and (4) show results for the fund-quarter regression in Equation 2.17. The quarterly sample is from 1980 to 2023. Standard errors (in parentheses) in columns (1) and (2) are two-way clustered by fund and stock-quarter, while standard errors in columns (3) and (4) are two-way clustered by fund and quarter. *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.



(A.) Mutual Funds Distribution to Investors.



(B.) Cash Dividends Payments by Firms to Mutual Funds.

FIGURE 2.6: Mutual Funds and Dividend Distribution by Calendar Month. This figure shows the share of total annual mutual funds' distribution to investors (Panel A) and cash dividends payments by firms (Panel B) occurring in each calendar month. In Panel A, we plot the fraction of total dollar distributions made by U.S. equity mutual funds to their investors by calendar month, using data from 1980 to 2023. Panel B plots the fraction of cash dividend payment by calendar month for all U.S. common stocks in our sample from 1980 to 2023. The dashed horizontal line represents the uniform benchmark if funds distributions (and cash dividends payment) were to follow a uniform distribution (i.e., $1/12 \approx 8.3\%$).

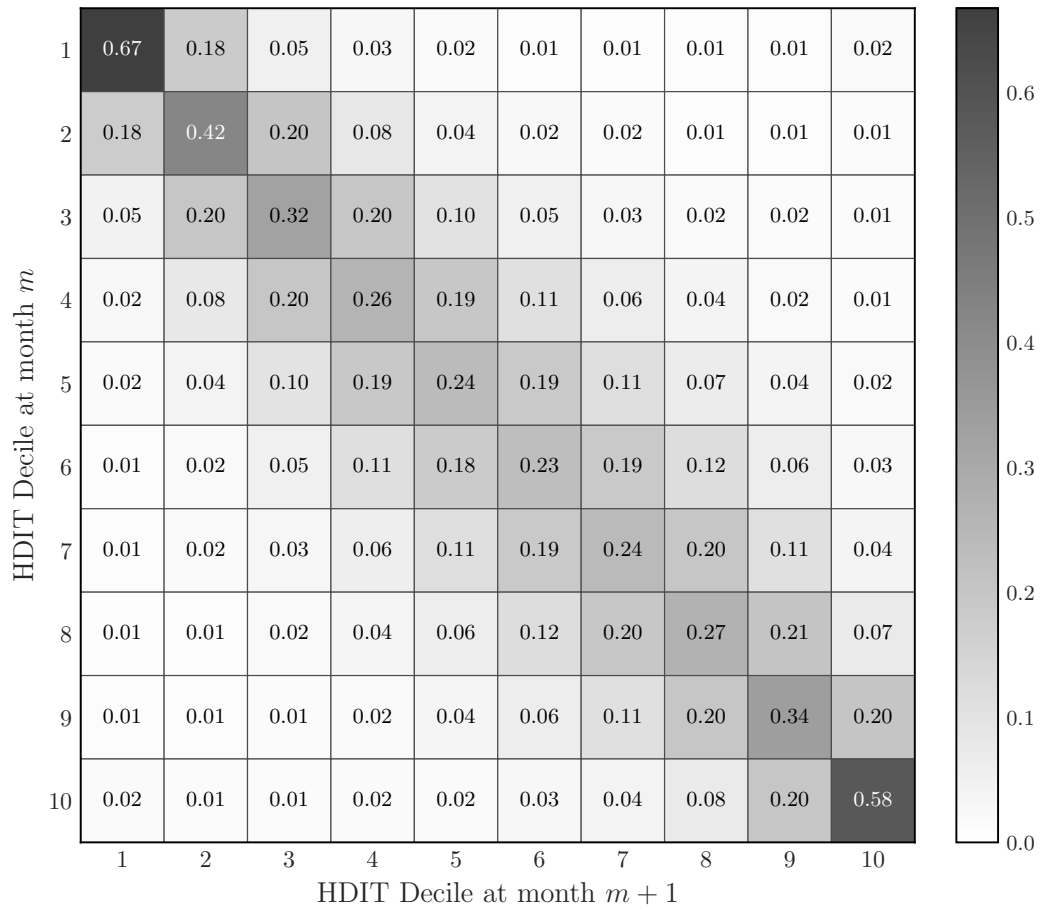


FIGURE 2.7: **Transition Matrix of HDIT by Decile.** This figure shows the probability of transitioning from HDIT decile i at month t (rows) to HDIT decile j at month $t + 1$ (columns). For each month from 1980 to 2023, we assign stocks to deciles based on the within-month distribution of aggregate HDIT, and compute the fraction of stock-month observations in decile i that move to decile j in the following month. Darker cells indicate higher transition probabilities.

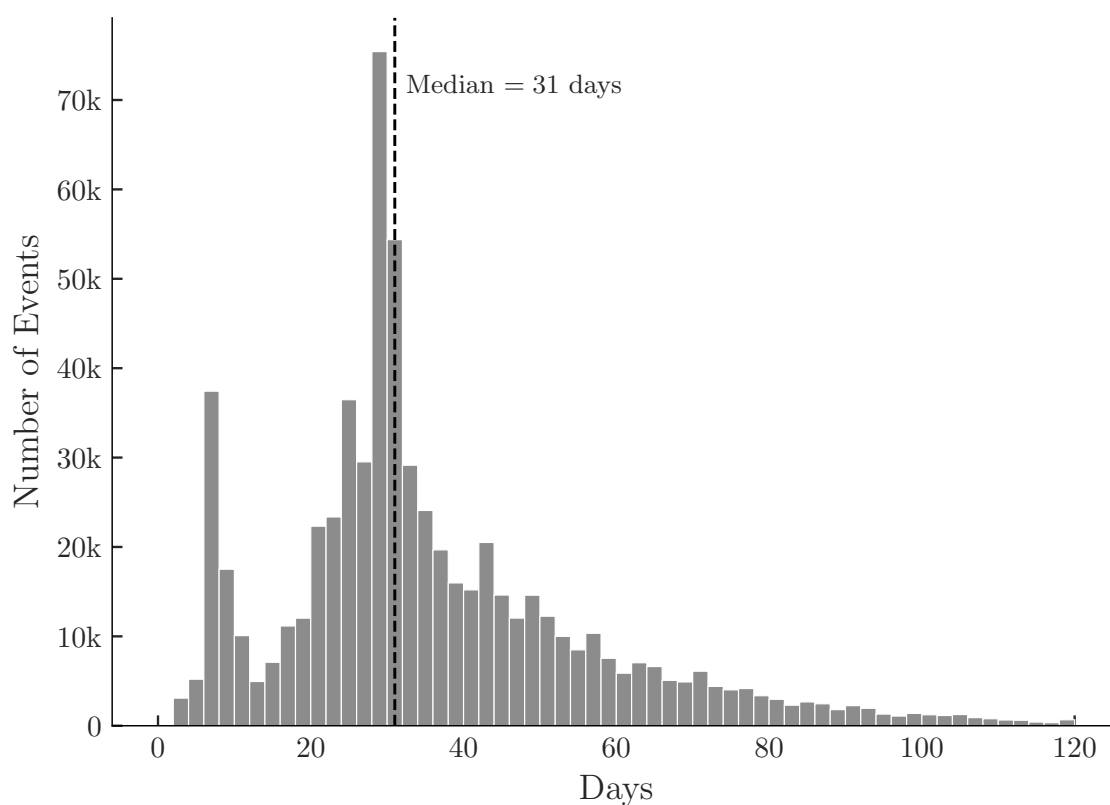
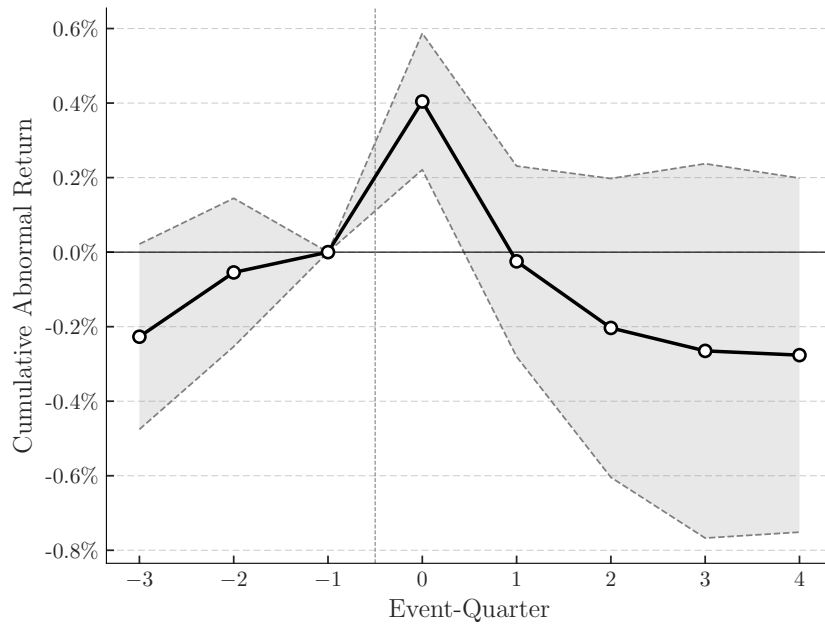


FIGURE 2.8: **Gap Between Dividend Announcement and Payment (in days).** This figure plots the distribution of the number of days between cash dividends' announcement date and the actual dividend payment date for all cash dividends paid in our sample from 1980 to 2023. The dashed vertical line indicates the median gap of 31 days (the mean is 39 days).

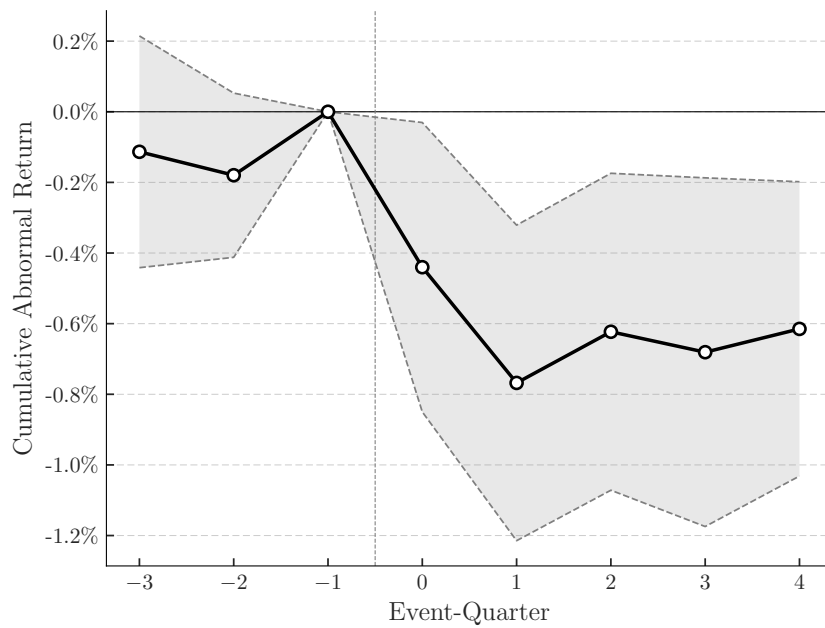
2.7.3 Appendix Figures



FIGURE 2.9: **Word Cloud of Data Technologies' Descriptions.** The figure shows the word cloud of data plugins' features in our full sample. For each data technology, we obtain 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.



(A.) Demand Pressure from Mutual Funds' Purchases.



(B.) Demand Pressure from Mutual Funds' Sales.

FIGURE 2.10: Mutual Funds Demand Pressure and Stock Returns.. This figure plots the quarterly average abnormal returns of stocks subject to mutual fund demand pressure from abnormal fund flows, following the approach of Dessaint et al. (2026). Panel A shows the effect of fire purchases driven by extreme fund inflows on stock returns. Panel B displays the effect of fire sales driven by extreme fund outflows on stock returns. The quarterly sample is from 1993 to 2023. The sample starts in 1993, as mutual fund net flows are reliably available in CRSP Mutual Funds only after March 1993 (Pástor et al., 2015).

2.7.4 Appendix Tables

Dependent variable:	CAPEX			TotalQ	TotalQ $\times$ DATA
	(1)	(2)	(3)	(4)	(5)
TotalQ _{<i>i,t-1</i>}	0.023*** (0.003)	0.028*** (0.004)	0.118*** (0.005)		
TotalQ _{<i>i,t-1</i>} $\times$ DATA _{<i>i,t-1</i>}	-0.010*** (0.002)	-0.014*** (0.005)	-0.038*** (0.010)		
DATA _{<i>i,t-1</i>}	0.063*** (0.015)	0.047*** (0.015)	-0.284*** (0.063)	0.079 (0.087)	0.336*** (0.035)
Size _{<i>i,t-1</i>}		-0.056*** (0.004)	-0.074*** (0.005)	-0.138*** (0.035)	-0.012** (0.005)
Cash Flow _{<i>i,t</i>}		0.026*** (0.003)	0.031*** (0.003)	0.204*** (0.016)	0.013*** (0.004)
HDIT _{<i>i,t-1</i>}				0.606*** (0.073)	0.014 (0.008)
HDIT _{<i>i,t-1</i>} $\times$ DATA _{<i>i,t-1</i>}				-0.087 (0.084)	0.169*** (0.019)
Estimator	2SLS	2SLS	OLS	First-stage	First-stage
Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.354	0.354	0.354	1.777	1.777
Outcome SE	0.460	0.460	0.460	3.786	3.786
Obs.	92,384	92,421	112,513	92,384	92,384
Adj. R ²	0.051	0.073	0.303	0.524	0.426
Kleibergen-Paap F-stat	16.425	16.144			

TABLE 2.16: **Robustness of Main Results using Total Q:** This table shows robustness tests for the main results using Total Q (Peters and Taylor, 2017). The dependent variable in columns (1) to (3) is CAPEX. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Total Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data tool in place in year t (DATA_{*i,t*}), and controls for firm-year characteristics. All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Dependent variable:	CAPX/Total Assets			
	(1)	(2)	(3)	(4)
$Q_{i,t-1}$	0.003*** (0.000)	0.005*** (0.001)	0.012*** (0.001)	0.011*** (0.001)
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.001*** (0.000)	-0.003*** (0.001)	-0.007*** (0.001)	-0.007*** (0.001)
$\text{DATA}_{i,t-1}$	0.019*** (0.004)	0.016*** (0.004)	0.009*** (0.002)	-0.028*** (0.008)
$\text{Size}_{i,t-1}$		-0.008*** (0.001)		-0.010*** (0.001)
Cash Flow $_{i,t}$		0.004*** (0.000)		0.004*** (0.001)
Estimator	2SLS	2SLS	OLS	OLS
Firm FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Outcome mean	0.060	0.060	0.060	0.060
Outcome SE	0.079	0.079	0.079	0.079
Obs.	100,205	101,629	128,216	122,905
Adj. R^2	0.041	0.061	0.528	0.545
Kleibergen-Paap F-stat	36.531	37.511		

TABLE 2.17: **Robustness using Different Measure of Investment:** This table shows robustness tests for the main results in the paper, using CAPX over Total Assets as measure of investment. The dependent variable is CAPX over Total Assets across all specifications. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Total Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) and (4) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data tool in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. All regressors are standardized by within-firm variation. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Dependent variable:	Intangible Investment					
	(1)	(2)	(3)	(4)	(5)	(6)
$Q_{i,t-1}$	0.007*** (0.000)	0.003*** (0.001)	0.003*** (0.001)			
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.002*** (0.001)	0.000 (0.000)	-0.001 (0.001)			
Total $Q_{i,t-1}$				0.005*** (0.000)	0.003*** (0.001)	0.004*** (0.001)
Total $Q_{i,t-1} \times \text{DATA}_{i,t-1}$				-0.001* (0.001)	0.000 (0.000)	-0.001 (0.001)
Estimator	OLS	2SLS	2SLS	OLS	2SLS	2SLS
Controls	✓	×	✓	✓	×	✓
Firm FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Outcome mean	0.221	0.221	0.221	0.221	0.221	0.221
Outcome SE	0.102	0.102	0.102	0.102	0.102	0.102
Obs.	55,676	47,808	46,634	55,423	47,580	46,418
Adj. R^2	0.601	0.015	0.046	0.610	0.035	0.065
Kleibergen-Paap F-stat		56.55	53.90		17.016	16.893

TABLE 2.18: **Placebo on R&D and Intangible Investment:** This table shows the results of panel regression in which the dependent variable is the one-year-ahead intangible investment, where intangible investment is defined following [Peters and Taylor \(2017\)](#) as $R\&D + 0.30 \times SG\&A$ divided by Intangible Capital. Column (1) and (4) show results for the OLS estimates, while columns (2), (3), (5), and (6) reports results for the instrumental variable (IV) estimates where Tobin's Q and TotalQ are instrumented by the Hypothetical Dividend Induced Trading (HDIT). The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics. The control variables (omitted for brevity) include a lagged firm's size ($\log$ assets) and contemporaneous cash flows over assets. All regressors are standardized by within-firm variation. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Dependent variable:	CAPEX			Q	Q × DATA
	(1)	(2)	(3)	(4)	(5)
$Q_{i,t-1}$	0.013*** (0.002)	0.017*** (0.002)	0.085*** (0.005)		
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.005*** (0.001)	-0.007*** (0.001)	-0.037*** (0.009)		
$\text{DATA}_{i,t-1}$	0.102*** (0.021)	0.096*** (0.019)	-0.112* (0.060)	0.085* (0.046)	1.110*** (0.046)
$\text{Size}_{i,t-1}$		-0.048*** (0.004)	-0.051*** (0.005)	-0.359*** (0.019)	-0.029*** (0.007)
Cash Flow $_{i,t}$		0.033*** (0.003)	0.032*** (0.002)	0.084*** (0.011)	0.007*** (0.002)
$\text{HDIT}_{i,t-1}$				0.566*** (0.031)	0.020*** (0.004)
$\text{HDIT}_{i,t-1} \times \text{DATA}_{i,t-1}$				-0.090** (0.037)	0.316*** (0.027)
Estimator	2SLS	2SLS	OLS	First-stage	First-stage
Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.353	0.353	0.353	1.985	1.985
Outcome SE	0.460	0.460	0.460	1.772	1.772
Obs.	100,206	100,206	121,055	100,206	100,206
Adj. R^2	0.037	0.056	0.286	0.611	0.814
Kleibergen-Paap F-stat	92.115	87.764			

TABLE 2.19: **Robustness of Main Results excluding Google Analytics:** This table shows robustness excluding Google Analytics from the set of data-related tools. The dependent variable is the one-year-ahead firm's investment. The dependent variable in columns (1) to (3) is CAPEX. Columns (1) and (2) reports results for the instrumental variable (IV) estimates, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (3) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$) excluding Google Analytics, and controls for firm-year characteristics. All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Instrumental Variable:	Fire Purchases		Fire Sales	
	(1)	(2)	(3)	(4)
$Q_{i,t-1}$	0.206 (0.134)	0.984 (7.427)	0.137** (0.061)	0.021 (0.084)
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$	-0.409** (0.165)	-1.628 (11.807)	-0.447*** (0.067)	-0.180 (0.129)
$\text{DATA}_{i,t-1}$	0.613** (0.254)	2.403 (17.310)	0.667*** (0.109)	0.276 (0.189)
$\text{Size}_{i,t-1}$		0.160 (1.665)		-0.060*** (0.021)
$\text{Cash Flow}_{i,t}$		-0.033 (0.512)		0.035*** (0.007)
Estimator	2SLS	2SLS	2SLS	2SLS
Firm FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Outcome mean	0.351	0.351	0.351	0.351
Outcome SE	0.455	0.455	0.455	0.455
Obs.	77,520	77,520	77,520	77,520
Adj. R^2	-0.198	-6.946	-0.246	-0.058
Kleibergen-Paap F-stat	0.853	0.008	3.142	1.303

TABLE 2.20: **Robustness of Main Results using Mutual Funds' Demand Pressure:** This table shows results of panel regression in which the dependent variable is the one-year-ahead firm's investment, and where Tobin's Q is instrumented using mutual fund demand pressure from abnormal fund flows, following Dessaint et al. (2026). Columns (1) and (2) use positive demand pressure from fire purchases driven by abnormal net inflows into mutual funds, while columns (3) and (4) use negative demand pressure from fire sales driven by abnormal net outflows. The control variables include a firm's size (log assets) in quarter t and contemporaneous cash flows over assets. The annual sample is from 1993 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

	$E_{i,t+1}/A_{i,t}$		$E_{i,t+3}/A_{i,t}$		$E_{i,t+5}/A_{i,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\log(\text{Mkt Cap}/A)_{i,t}$	0.017*** (0.003)	0.016*** (0.003)	0.007** (0.004)	0.005 (0.004)	0.019*** (0.003)	0.017*** (0.003)
$\log(\text{Mkt Cap}/A)_{i,t} \times \text{DATA}_{i,t}$	-0.005** (0.002)	-0.007** (0.003)	-0.003 (0.004)	-0.002 (0.004)	-0.001 (0.004)	0.006 (0.005)
Controls	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Time FE	✓	×	✓	×	✓	×
Industry×Time FE	×	✓	×	✓	×	✓
Outcome mean	0.028	0.028	0.055	0.055	0.084	0.084
Outcome SE	0.222	0.222	0.267	0.267	0.314	0.314
Obs.	142,111	140,199	116,962	115,400	96,767	95,467
Adj. R^2	0.814	0.817	0.627	0.635	0.562	0.571

TABLE 2.21: **Price Informativeness and Data Technologies (Bai et al., 2016)**: This table shows results of panel regression in which the dependent variable is the h -years-ahead firm's earnings over current assets, for $h = \{1, 3, 5\}$. The regressors are a the $\log$ market capitalization over assets, and an interaction term with a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$), and controls for firm-year characteristics (omitted for brevity). The coefficient on the $\log$ market capitalization over assets captures price informativeness (Bai et al., 2016). The control variables include a firm's current earnings over assets, size ($\log$ assets), and cash flows over assets in quarter t . Industry×Time fixed effects use two-digits SIC industry codes. The annual sample is from 1980 to 2023. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

Dependent variable:	Jump Ratio (Weller, 2018)		CAPEX		
	(1)	(2)	(3)	(4)	(5)
$Q_{i,t-1}$		0.001 (0.002)	0.022*** (0.005)	0.026*** (0.006)	0.096*** (0.007)
$Q_{i,t-1} \times \text{DATA}_{i,t-1}$			-0.012*** (0.004)	-0.014** (0.007)	-0.043*** (0.010)
$\text{DATA}_{i,t-1}$	-0.263*** (0.011)	-0.276*** (0.011)	0.129*** (0.034)	0.117*** (0.033)	-0.037 (0.079)
$\text{Size}_{i,t-1}$		0.039*** (0.006)		-0.039*** (0.007)	-0.040*** (0.007)
$\text{Cash Flow}_{i,t}$		0.090*** (0.015)		0.026*** (0.005)	0.025*** (0.004)
Estimator	Stacked DiD		2SLS	2SLS	OLS
Cohort×Firm FE	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓
Outcome mean	0.361	0.361	0.368	0.368	0.368
Outcome SE	0.336	0.336	0.478	0.478	0.478
Obs.	146,301	141,969	152,834	152,834	176,597
Adj. R^2	0.198	0.208	0.059	0.070	0.270
Kleibergen-Paap F-stat			50.996	54.582	

TABLE 2.22: **Robustness of Main Results on subsample with positive effect on Price Informativeness:** This table shows robustness of the main results for a subsample in which the adoption of data-related tools is associated with an increase in measures of price informativeness Weller (2018). We identify the subsample for this robustness using influence functions for the effect of data-related tools on the Weller (2018) jump ratio. We compute the influence of each firm on the estimates (DFBETA) and keep only firms above the median, i.e., we isolate firms for which the effect of data-related tools on price informativeness is positive. Columns (1)-(2) show the effect of data technology on stock price jump ratio around earning announcements (Weller, 2018). A higher jump ratio is associated with less informative stock prices. Columns (3) and (4) reports results for the instrumental variable (IV) estimates in the subsample, where Tobin's Q is instrumented by the Hypothetical Dividend Induced Trading (HDIT). Column (5) reports the OLS results. The regressors are a dummy equal to one if a firm i has a data technology in place in year t ($\text{DATA}_{i,t}$) excluding Google Analytics, and controls for firm-year characteristics. All estimates are in the stacked sample (Gormley and Matsa, 2011). All regressors are standardized by within-firm variation. Columns (4) and (5) report the first-stage results. The annual sample is from 1980 to 2023. All columns include firm and year fixed-effects. All standard errors are two-way clustered by firm and year (in parentheses). *, **, and *** denote statistical significance at the 10%, 5% and 1% respectively.

References

- Abis, Simona (2022) "Man vs. Machine: Quantitative and Discretionary Equity Management." Working Paper, University of Colorado at Boulder.
- Alexander, Lewis and Janice Eberly (2018) "Investment Hollowing Out," *IMF Economic Review*, Vol. 66, No. 1, pp. 5–30.
- Ali, Ashiq, Kelsey Wei, and Yibin Zhou (2011) "Insider Trading and Option Grant Timing in Response to Fire Sales (and Purchases) of Stocks by Mutual Funds," *Journal of Accounting Research*, Vol. 49, No. 3, pp. 595–632.
- Amazon (2025) "Privacy Cookie Policy." Amazon.
- Babina, Tania (2026) "Understanding Firms' AI Efforts and Their Economic Impact," Working Paper 35123, National Bureau of Economic Research.
- Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2023) *Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition*, pp. 75–117: University of Chicago Press.
- (2024a) "Artificial Intelligence and Firms' Systematic Risk." Working Paper, University of Maryland.
- (2024b) "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, Vol. 151, p. 103745.
- Bai, Jennie, Thomas Philippon, and Alexi Savov (2016) "Have financial markets become more informative?" *Journal of Financial Economics*, Vol. 122, pp. 625–654.
- Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang (2022) "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics*, Vol. 144, No. 2, pp. 370–395.
- Baker, Malcolm, Jeremy Stein, and Jeffrey Wurgler (2003) "When Does the Market Matter? Stock Prices and the Investment of Equity-Dependent Firms," *Quarterly Journal of Economics*, Vol. 3, No. 118, pp. 969–1005.
- Baley, Isaac and Laura Veldkamp (2025) *The Data Economy: Tools and Applications*: Princeton University Press.
- Blankespoor, Elizabeth, Bradley Hendricks, Joseph Piotroski, and Christina Synn (2022) "Real-time revenue and firm disclosure," *Review of Accounting Studies*, Vol. 27, p. 1079–1116.
- Bond, Philip, Alex Edmans, and Itay Goldstein (2012) "The real effects of financial markets," *Annual Reviews of Financial Economics*, Vol. 4, pp. 339–360.
- Bonelli, Maxime (2025) "Data-Driven Investors," *The Review of Financial Studies*, p. hhaf078, 10.
- Bonelli, Maxime and Thierry Foucault (2025) "Does Big Data Devalue Traditional Expertise? Evidence from Active Funds." Working Paper, HEC Paris.
- Brynjolfsson, Erik and Kristina McElheran (2016) "The Rapid Adoption of Data-Driven Decision-Making," *The American Economic Review*, Vol. 106, No. 5, pp. 133–139.

- Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond (2023) "Generative AI at Work," Working Paper 31161, National Bureau of Economic Research.
- Call, Andrew C., Paul Hribar, Douglas J. Skinner, and David Volant (2024) "Corporate managers' perspectives on forward-looking guidance: Survey evidence," *Journal of Accounting and Economics*.
- Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L Zhang (2023) "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI," *The Review of Financial Studies*, Vol. 36, No. 9, pp. 3603–3642, 03.
- Chen, Huaizhi (2024) "Cash Induced Demand," *Journal of Financial and Quantitative Analysis*, No. 53, pp. 195–220.
- Chen, Qi, Itay Goldstein, and Wei Jiang (2007) "Price Informativeness and Investment Sensitivity to Stock Prices," *Review of Financial Studies*, Vol. 20, No. 3, pp. 619–650.
- Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng (2025) "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts," *Management Science*, Vol. 71, No. 6, pp. 4599–4621.
- Ciconte, Warren, Marcus Kirk, and Jennifer W. Tucker (2014) "Does the midpoint of range earnings forecasts represent managers' expectations?" *Review of Accounting Studies*, Vol. 19, pp. 628–660.
- Cockburn, Iain M, Rebecca Henderson, and Scott Stern (2018) "The Impact of Artificial Intelligence on Innovation," Working Paper 24449, National Bureau of Economic Research.
- Cohen, Lauren, Christopher Malloy, and Lukasz Pomorski (2012) "Decoding Inside Information," *The Journal of Finance*, Vol. 67, No. 3, pp. 1009–1043.
- Coller, Maribeth and Teri L. Yohn (1997) "Management forecasts and information asymmetry: an examination of bid-ask spreads," *Journal of Accounting Research*, Vol. 35, No. 2, pp. 181–191.
- Comin, Diego and Bart Hobijn (2010) "An Exploration of Technology Diffusion," *American Economic Review*, Vol. 100, No. 5, pp. 2031–2059.
- Cook, R. Dennis and Sanford Weisberg (1982) "Residuals and Influence in Regression," *Biometrical Journal*, Vol. 27, No. 1.
- Davila, Eduardo and Cecilia Parlatore (2025) "Identifying price informativeness." Working paper.
- De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020) "The Rise of Market Power and the Macroeconomic Implications*," *Quarterly Journal of Economics*, Vol. 135, No. 2, pp. 561–644.
- Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2024) "Does Alternative Data Improve Financial Forecasting? The Horizon Effect." *Journal of Finance*, Forthcoming.
- Dessaint, Olivier, Thierry Foucault, Laurent Frésard, and Adrien Matray (2019) "Noisy Stock Prices and Corporate Investment," *Review of Financial Studies*, Vol. 32, No. 7, pp. 2625–2672.
- Dessaint, Olivier, Naveen Gondhi, and Joel Peress (2026) "Learning about Discount Rates." Working Paper, INSEAD.

- Diamond, Douglas W. (1985) "Optimal Release of Information by Firms," *The Journal of Finance*, Vol. 40, No. 4, pp. 1071–1094.
- Dou, Winston Wei, Itay Goldstein, and Yan Ji (2025) "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency," Working Paper 34054, National Bureau of Economic Research.
- Dugast, Jerome and Thierry Foucault (2025) "Equilibrium Data Mining and Data Abundance," *The Journal of Finance*, Vol. 80, No. 1, pp. 211–258.
- Edmans, Alex, Itay Goldtsein, and Wei Jiang (2012) "The Real Effects of Financial Markets: The Impact of Prices on Takeovers," *Journal of Finance*, Vol. 67, No. 3, pp. 933–971.
- Eeckhout, Jan and Laura Veldkamp (2022) "Data and Markups: A Macro-Finance Perspective," NBER Working Paper 30022.
- Erickson, Timothy and Toni Whited (2000) "Measurement Error and the Relationship between Investment and q ," *Journal of Political Economy*, Vol. 108, No. 5, pp. 1027–1057.
- Ferracuti, Elia, Minjae Koo, Mary Lee, and Stephen Stubben (2024) "Acquisition of Customer Information and Corporate Decision Making." Working Paper, Hong Kong University Business School.
- Foucault, Thierry and Laurent Fresard (2014) "Learning from peers' stock prices and corporate investment," *Journal of Financial Economics*, Vol. 111, No. 3, pp. 554–577.
- Foucault, Thierry, Leonardo Gambacorta, Wei Jiang, and Xavier Vives (2025) "Barcelona 7: Artificial Intelligence in Finance." CEPR Press, Paris & London.
- Frésard, Laurent, Gerard Hoberg, and Gordon M Phillips (2019) "Innovation Activities and Integration through Vertical Acquisitions," *The Review of Financial Studies*, Vol. 33, No. 7, pp. 2937–2976, 09.
- 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.
- Goldstein, Itay (2023) "Information in Financial Markets and Its Real Effects," *Review of Finance*, Vol. 27, pp. 1–32, 08.
- Goodman-Bacon, Andrew (2021) "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, Vol. 225, No. 2, pp. 254–277.
- Gormley, Todd A. and David A. Matsa (2011) "Growing Out of Trouble? Corporate Responses to Liability Risk," *Review of Financial Studies*, Vol. 24, No. 8, pp. 2781–2821.
- Gormsen, Niels Joachim and Kilian Huber (2025) "Corporate Discount Rates," *American Economic Review*, Vol. 115, No. 6, p. 2001–49.
- Gutiérrez, G. and T. Philippon (2017) "Investment-less Growth: An Empirical Investigation," *Brookings Papers on Economic Activity*.

- Hayashi, F. (1982) "Tobin's marginal q and average q: A neoclassical interpretation," *Econometrica*, Vol. 50, pp. 215–224.
- Hoberg, Gerard and Vojislav Maksimovic (2015) "Redefining Financial Constraints: A Text-Based Analysis," *Review of Financial Studies*, Vol. 28, No. 5, pp. 1312–1352.
- Khan, Mozaffar, Leonid Kogan, and George Serafaim (2012) "Mutual Fund Trading Pressure: Firm-Level Stock Price Impact and Timing of SEOs," *The Journal of Finance*, Vol. 67, No. 4, pp. 1371–1395.
- Labro, Eva, Mark Lang, and James D. Omartian (2023) "Predictive analytics and centralization of authority," *Journal of Accounting and Economics*, Vol. 75, No. 1, p. 101526.
- Linn, Matthew and Daniel Weagley (2023) "Uncovering Financial Constraints," *Journal of Financial and Quantitative Analysis*, p. 1–36.
- Liu, Zack and Adam Winegar (2025) "Economic magnitudes within reason," *Journal of Corporate Finance*, Vol. 91.
- Lou, Dong (2012) "A Flow-Based Explanation for Return Predictability," *Review of Financial Studies*, Vol. 25, No. 12, pp. 3457–3489, 12.
- Lu, Yao, Gordon Phillips, and Jia Yang (2025) "The Impact of Cloud Computing and AI on Industry Concentration and Dynamics." Working Paper, Tuck School of Business.
- 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.
- Morck, Randall, Andrei Shleifer, and Robert Vishny (1990) "The Stock Market and Investment: Is the Market a Sideshow?" *Brookings Papers on Economic Activity*, Vol. 21, pp. 157–216.
- Morris, Stephen and Hyun Song Shin (2002) "Social Value of Public Information," *American Economic Review*, Vol. 92, No. 5, p. 1521–1534, December.
- Nickell, Stephen J. (1996) "Competition and Corporate Performance," *Journal of Finance*, Vol. 104, No. 4, pp. 724–746.
- Nike (2025) "Privacy Cookie Policy." Nike.
- Pástor, Luboš, Robert F. Stambaugh, and Lucian A. Taylor (2015) "Scale and Skill in Active Management," *Journal of Financial Economics*, Vol. 116, No. 1, pp. 23–45.
- Peters, Ryan and Lucian Taylor (2017) "Intangible capital and the investment-q relation," *Journal of Financial Economics*, Vol. 100, pp. 251–272.
- Schmickler, Simon and Pedro Tremacoldi-Rossi (2023) "Spillover Effects of Payouts on Asset Prices and Real Investment." Working paper.
- Schmidheiny, Kurt and Sebastian Sieglöcher (2023) "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization," *Journal of Applied Econometrics*, Vol. 38, No. 5, pp. 695–713.
- Subrahmanyam, Avanidhar and Sheridan Titman (1999) "The Going Public Decision and the Development of Financial Markets," *Journal of Finance*, Vol. 54, No. 3, pp. 1045–82.

- van der Beck, Philippe (2024) "Flow-Driven ESG Returns," *Working paper*.
- Verrecchia, Robert E. (1982) "The Use of Mathematical Models in Financial Accounting," *Journal of Accounting Research*, Vol. 20, No. Supplement, pp. 1–42.
- Wardlaw, Malcom (2020) "Measuring Mutual Fund Flow Pressure as Shock to Stock Returns," *Journal of Finance*, Vol. 75, No. 6, pp. 3221–3243.
- Weller, Brian M. (2018) "Does Algorithmic Trading Reduce Information Acquisition?" *The Review of Financial Studies*, Vol. 31, No. 6, pp. 2184–2226.
- Yan, Dong (2024) "Do private firms (mis)learn from the stock market?" *Review of Finance*, Vol. 28, No. 5, pp. 1483–1511, 07.
- Zanotti, Marco (2025) "Asset (and Data) Managers." Working Paper, Swiss Finance Institute.

3 Does a Firm's Data Make its Price More Informative?

3.1 Introduction

One of the most prominent trends in the modern economy is firms' adoption of new information technologies, such as artificial intelligence and data analytics. Across industries and countries, firms increasingly collect and process large volumes of information that did not exist decades ago, and use it to guide decisions from product development to pricing. What a firm knows also determines what information it can disclose publicly, and thereby what financial markets can learn. Managers decide what information to disclose, while investors devote substantial resources to producing information of their own, and the incentives of each depend on the information produced by the other. Thus, the consequences of firms' technological adoption extend well beyond their internal decisions, spilling over into investors' incentives and, ultimately, into what stock prices reveal. Understanding how these technologies change the information produced by financial markets requires studying the two sides jointly, rather than each side in isolation.

In this paper, I ask how technologies that improve firms' internal information affect the information contained in the stock price. I develop a model to study the interaction between firms' optimal disclosure and investors' endogenous information acquisition. The firm's manager observes a private signal (a data point) that helps forecast the firm's uncertain future fundamental. She chooses how precisely to disclose her information to the public, trading off the benefit of a lower cost of capital (higher stock price) against the proprietary cost of public disclosure. At the same time, investors choose the precision of their private information by searching for profitable signals in a pool of raw data. The search cost rises endogenously with the firm's public disclosure, since disclosure reveals information to the public and the stock price incorporates that information. Thus, investors cannot profit from those data points and must filter them out when searching for profitable signals, imposing a negative externality on their cost of acquiring information. The joint equilibrium of disclosure

and information acquisition determines: (i) the firm's optimal disclosure, and (ii) the informativeness of the stock price. The model delivers two main results. First, technologies that increase the firm's internal information unambiguously increase the manager's optimal disclosure. Second, those technologies can nonetheless reduce the informativeness of the stock price—defined as the extent to which prices predict the realization of future fundamentals—because disclosure raises investors' cost of producing information of their own.

The paper's central departure from standard models of disclosure is that investors' cost of acquiring information is not fixed but rises endogenously with the precision of the firm's disclosure. This departure creates two distinct mechanisms through which public disclosure crowds out private information. The first mechanism is common to standard models (see, for example, [Verrecchia, 1982](#); [Diamond, 1985](#); [Gao and Liang, 2013](#); [Goldstein and Yang, 2017a](#)). In these settings, disclosure crowds out private information by reducing its marginal *benefit* for investors. A more informative price leaves investors less to gain from a given private signal. The second mechanism, which is novel, acts through an increase in the marginal *cost* of searching for profitable private information. When the firm discloses more, the total volume of information available about the firm grows and the stock price promptly incorporates the additional information disclosed. Yet, the stock of signals that remain profitable for trading does not grow proportionally with public disclosure—it remains unchanged, or it may actually deteriorate. Investors must filter their signals through an expanding pool of (unprofitable) data points, and finding a profitable signal becomes more expensive. The second mechanism is crucial for the paper's main result. When disclosure reduces only the benefit of private information, crowding-out is never strong enough to overturn the direct gain in public precision, and price informativeness always rises when the firm becomes better informed. With the additional “marginal cost” mechanism I introduce in this paper, however, this conclusion can reverse. When public disclosure becomes sufficiently precise, investors' search cost rises so much that the loss of private information exceeds the gain in public precision, and price informativeness falls.

The model delivers an exact condition for when price informativeness falls. The condition combines two forces. First, investors' search cost of finding a profitable signal must rise steeply as public information accumulates. In other words, each additional management disclosure makes profitable signals substantially harder to find within the growing volume of data available about the firm. For example, this can arise because media outlets, commentators, analysts, or social media pick up the additional information disclosed by the firm, and generate derivative content

based on the firm's disclosure. Second, the stock price must rely heavily on the information produced by investors. When most of what the price reveals comes from investors' private research, diminishing their incentives to search destroys much of the price's content. Together, the two conditions identify when price informativeness deteriorates with a firm's internal information. Moreover, since price informativeness always increases in traditional models with constant cost of information acquisition, any observed decline traces back to the rising cost of private search. This prediction contrasts with the evidence that *investors'* access to more data has made prices more informative (??). Here, I study what happens when the *firm* has access to more data, and the impact on price informativeness can reverse.

The additional crowding-out mechanism also generates spillovers across firms. I extend the model to a two-firm economy in which investors allocate scarce research capacity between the two firms, so that devoting more effort to one raises the marginal cost of researching the other. Suppose firm A adopts a new data technology—say, customer analytics or AI—that improves the precision of its internal information, while firm B does not. Firm A will therefore (optimally) disclose more precise information to the public, which makes searching for profitable signals about A less valuable to investors and, relative to firm B, more expensive. Thus, investors respond by reallocating research effort toward firm B. As a result, firm B's stock price becomes more informative and its cost of capital falls, even though nothing about firm B's own information has changed. Notably, the two firms' fundamentals are independent, so the spillover does not originate from correlated cash flows or product-market connections. This result comes entirely from the reallocation of investors' research effort. Technology adoption thus creates an informational externality, and part of its market impact shows up not at the adopting firm but at its peers.

A key component of the model is the investors' endogenous search cost that increases with the firm's public disclosure. This cost function is not imposed or assumed, but I derive it from a model of sequential search and content creation. Investors sample data points one by one from the total pool of raw datasets available about the firm, but only some of those data—say, satellite images or web-traffic information—carry profitable information beyond what the firm has disclosed (and therefore, beyond what the price already contains). The rest is “derivative content”, produced by a competitive content creation sector (e.g., media, analysts, data vendors, social media) that uses the firm's disclosure as a free input and enters the industry as long as investors' attention makes content profitable. More disclosure attracts more derivative content, shrinks the share of profitable data points in the pool of raw data, and thus forces investors to search longer. The resulting cost of searching for

a profitable data point is exactly the search cost function of the main model. It is worth noticing that this search cost technology is one formulation of a more general economic force. A limited-attention framework would deliver qualitatively the same economic predictions. When investors have finite capacity to process information, a growing pool of public data and derivative content absorbs attention that could otherwise be used to generate private information (Hirshleifer and Teoh, 2003; Peng and Xiong, 2006).

A natural concern might be that sophisticated investors do not search at random in the pool of raw datasets. For example, a hedge fund manager may recognize that satellite data are more likely to contain novel information than an analyst report commenting on the firm's guidance. However, building on the partially directed search framework in Menzio (2007), I extend the model and show that the main mechanism holds when investors have insight on which data points might be more profitable than others. Even when investors can screen data sources with high accuracy—for example, using AI to understand which types of information are incorporated into prices—imperfect screening lets a fraction of redundant content through, and as the pool grows, these false positives eventually overwhelm the fixed stock of profitable signals. An accurate screening technology thus attenuates the externality of public disclosure and delays the decline of price informativeness, but it does not prevent it. Moreover, I show that even sophisticated screening technologies become nearly useless as the amount of data created by content creators expands.

Is the negative externality of public disclosure economically large? Although investors' information sets are not directly observable, recent evidence suggests that the cost of extracting profitable information from available data is sizable. de Haan et al. (2025) build an AI-based analyst that processes only publicly available data (e.g., corporate disclosures, news, transcripts, and structured databases) and estimate its information-processing cost at roughly \$17 million per quarter per fund, which is more than five times the average fund's revenue from alpha or from management fees. In the model's language, the premium for finding profitable signals in the pool of raw data is large. The condition for a negative effect on price informativeness, on the other hand, requires surprisingly little. In a calibrated example, price informativeness declines once each signal disclosed by the firm generates about 1.15 to 1.25 units of derivative content. For instance, two news articles commenting on an earnings release would be enough for price informativeness to decline. Given the explosion of content created around each corporate disclosure, these magnitudes appear well within plausible ranges.

The results have implications for disclosure regulation in a data economy. In

the model, the manager chooses disclosure to maximize the stock price and does not internalize the effect of her disclosure on investors' cost of searching for profitable information. Privately optimal disclosure can therefore exceed the level that maximizes the information available to the market. Mandatory disclosure standards can widen this wedge, since forcing firms to generate public information inflates the pool of data around them and can reduce, rather than increase, the total information available to the market. The cross-firm spillover, however, cuts in the opposite direction. Mandating disclosure at one firm redirects investors' research toward other firms, improving the overall information environment around firms with less precise internal information and lowering their cost of capital. These two forces push disclosure policy in opposite directions. The optimal design depends on their relative magnitudes, and different policy mandates can accommodate the trade-off between the informational cost at the disclosing firm and the spillover benefit to its peers.¹

The model is deliberately simple, with the minimum ingredients necessary to generate the results. The precision of the firm's internal information is exogenous. I do not model the firm's decision to adopt a data technology, but rather study the effect of adoption on the interaction between firms and financial markets once it occurs. Similarly, the manager does not learn from the stock price when making real decisions, so the model abstracts from the feedback effects studied in [Bond et al. \(2012\)](#) and [Goldstein and Yang \(2017a\)](#). These choices isolate the mechanism this paper focuses on: investors' production of information.

Taken together, these results caution against the common intuition that better informed firms always lead to more informative financial markets. The paper's contribution is a tractable framework in which the firm's internal information, its disclosure, and investors' search for profitable information are jointly determined. In such a framework the cost of finding profitable signals rises with the volume of public data. This endogenous cost of acquiring profitable data—microfounded by a content creation industry and sequential search, and plausible in magnitude—is enough to reverse the standard prediction that firms' access to more data always improves price informativeness, and to generate spillovers of data adoption across firms. More broadly, this framework shows that the informational consequences of firms' technology adoption cannot be understood from firms' decisions alone, but depend on how financial markets reorganize their own information production in response.

Related literature. The paper contributes to three main strands of research. First, the paper contributes to the literature on information acquisition in financial markets

¹Characterizing the optimal disclosure mandate that balances the two forces in different economies is a natural question for future work.

and price informativeness. This literature studies the trade-off between informative prices and costly information acquisition decisions (Grossman and Stiglitz, 1980; Hellwig, 1980; Verrecchia, 1982). Verrecchia (1982) analyzes both the extensive and intensive margins of information acquisition. In this paper, I use the intensive margin and endogenize the cost of information acquisition. Recent research studies how improvements in data processing affect investors' choice between learning about fundamentals or about others' demand (Farboodi and Veldkamp, 2020), trading and asset prices (Dugast and Foucault, 2018; Abis, 2022; Huang et al., 2022; Bonelli and Foucault, 2023), and market efficiency (Martin and Nagel, 2022). While most of these works focus on how investors' data availability affects information acquisition and price informativeness, I show that results can be starkly different when *firms*, rather than investors, have access to more informative data. My framework is inspired by Dugast and Foucault (2025), but differs in two respects: (i) the data abundance is endogenous (generated by the firm's own disclosure), and (ii) price informativeness can decrease, because the firm's disclosure creates a negative externality on investors' research costs.

Second, the paper contributes to the literature on the impact of technology on corporate disclosure.² Building on classic models of corporate disclosure, this literature asks whether (and how) new information technologies affect firms' optimal disclosure. Cao, Ma, Tucker and Wan (2018) show that firms facing stronger technological competition reduce product-related disclosures. This is consistent with proprietary disclosure costs rising with competitors' ability to process information. In general, product-market competitors' ability to imitate disclosed strategies deters disclosure (Cao, Du, Yang and Zhang, 2021). Cao, Jiang, Yang and Zhang (2023) show that easier machine readership of corporate filings changes how managers write disclosures, and recently Blankespoor, de Haan and Li (2026) find that a significant fraction of new corporate disclosure is now generated by Generative AI. Relative to this work, I study how the firm's internal data change its optimal disclosure and how the market's information production responds in equilibrium. Because the manager cares about the price level (or about the firm's cost of capital), she does not internalize the effect of her disclosure on investors' search costs, and price informativeness can fall.

Third, the paper contributes to the growing literature on the effects of new technologies in finance. This literature shows how technology affects financial forecasting (e.g., Chi, Hwang and Zheng, 2025; ?; ?; Dessaint, Foucault and Frésard,

²The main theoretical frameworks in this literature are Verrecchia (1983, 2001); Dye (1985); Diamond (1985); Diamond and Verrecchia (1991); Admati and Pfleiderer (2000), and Goldstein and Yang (2017b). Goldstein and Yang (2017a) review part of this literature and provide a general framework for understanding how public disclosure affects market quality, information production, and real efficiency.

2024; Cao, Jiang, Yang and Zhang, 2024), firms' production function (e.g., ?; Farboodi and Veldkamp, 2023, 2025; Abis and Veldkamp, 2024; Cong, Xie and Zhang, 2021) and capital allocation (Abis, 2022; Bonelli, 2025; ?; Bonelli and Foucault, 2023; Sheng, Sun, Yang and Zhang, 2025; Zanolini, 2026). This literature mostly studies how technology changes either *investors'* or *firms'* behavior separately. I study how firms' data availability changes the *interaction* between firms and financial markets. In doing so, this paper studies the equilibrium implications of data on corporate disclosure, information production, and prices.

3.2 Model

The model builds on a standard noisy rational-expectations equilibrium (NREE) (e.g., Hellwig, 1980; Verrecchia, 1982). A firm's stock has an uncertain liquidation value. The firm's manager observes a private signal about the fundamental and chooses how precisely to disclose it to a competitive stock market. Investors choose the precision of their private research at a cost that depends endogenously on the information environment around the firm. When firms disclose more public information, investors have more information that they can process and can learn from. However, public information gets impounded into the stock price and investors need more effort to search for profitable data points. This endogenous cost to generate informative signals about the firm is the main departure from a standard NREE setting.

3.2.1 The Firm

There are three dates, $t = 0$, $t = 1$, and $t = 2$. There is a unique firm in the economy, whose fundamental $\tilde{\theta} \sim N(0, \tau_\theta^{-1})$ is realized in $t = 2$, with $\tau_\theta > 0$. The manager of the firm is risk-neutral and has a private signal about the fundamental

$$s = \tilde{\theta} + \chi, \quad \chi \sim N(0, \tau_\chi^{-1}), \quad \chi \perp \tilde{\theta}, \quad (3.1)$$

where $\tau_\chi \geq 0$ is the precision of the firm's internal information. For example, managers can have private information about their operations, about the quality of their products, or about future product demand. A firm with more precise information about the future fundamental has higher τ_χ .

The manager can decide to disclose part of her information to the public. Formally, the manager can disclose a signal ω that consists in her private signal plus random

noise v :

$$\omega = s + v = (\tilde{\theta} + \chi) + v, \quad v \sim N(0, \tau_v^{-1}), \quad (3.2)$$

where $v \perp (\tilde{\theta}, \chi)$. The precision of disclosure noise (τ_v) governs the accuracy of the information that the manager communicates to the public. For instance, a manager who wants to disclose a substantial part of her internal information sets a high τ_v . On the other hand, when the manager prefers not to disclose part of her private information she can set $\tau_v \rightarrow 0$. The total precision of ω about the uncertain fundamental $\tilde{\theta}$ is a function of the manager's disclosure precision *and* her internal information:

$$\tau_\omega = \frac{\tau_\chi \tau_v}{\tau_\chi + \tau_v}. \quad (3.3)$$

The firm's internal information and its disclosure jointly determine how much the public can learn. Importantly, the firm's internal information and public disclosure are complements: accurate public disclosure is more informative when the manager has better knowledge about the future fundamental. I define the total public precision as

$$\tau_{pub} := \tau_\theta + \tau_\omega. \quad (3.4)$$

The manager represents current shareholders and in $t = 0$ maximizes the expected stock price net of the proprietary cost of disclosure:

$$U_M = \mathbb{E}[\tilde{P}] - C(\tau_v). \quad (3.5)$$

The disclosure cost, $C(\tau_v)$, might reflect the cost of revealing information to product-market rivals, the litigation risk from imprecise forecasts, or the direct compliance cost of producing disclosures. I assume $C(\tau_v)$ is convex in the precision of public disclosure ($C'(\tau_v) > 0$, $C''(\tau_v) > 0$). In other words, disclosing more precise information is costly, and becomes increasingly so as precision rises. Figure 3.1 summarizes the timing of the model.

3.2.2 The Stock Market

There is a unit mass of investors, indexed by i , with CARA utility $U(w) = -\exp(-\rho w)$, where w is the investor's terminal wealth in $t = 2$, and $\rho > 0$ is risk aversion. Investors can trade in two assets: a risk-free asset and a risky asset. The risk-free asset has a constant value of 1 and is in unlimited supply. The risky asset is usually thought of as equity issued by the firm, and pays the firm's uncertain cash flow $\tilde{\theta}$ at the final date $t = 2$. The risky asset is in fixed supply Q and investors trade competitively at an endogenous price $\tilde{P}$.

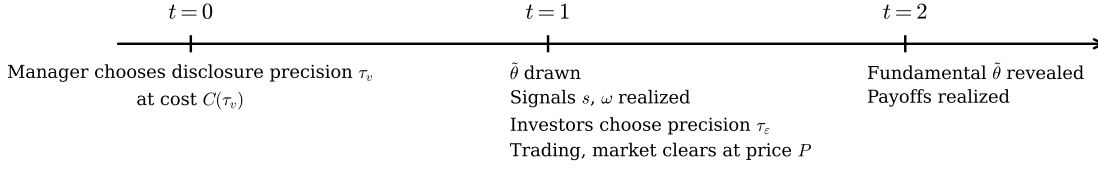


FIGURE 3.1: **Timeline of the Model.** This figure summarizes the timing of the model. At $t = 0$, the manager chooses the precision of public disclosure τ_v at cost $C(\tau_v)$. At $t = 1$, the fundamental $\tilde{\theta}$ is drawn, the signals s and ω are realized, investors choose the precision of private research τ_ε , and the market clears at price $\tilde{P}$. At $t = 2$, the fundamental $\tilde{\theta}$ is revealed and payoffs are realized.

In $t = 1$, each investor observes a private signal

$$y_i = \tilde{\theta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \tau_{\varepsilon_i}^{-1}), \quad (3.6)$$

with ε_i i.i.d. and $\varepsilon_i \perp (\tilde{\theta}, \chi, v)$. The precision of each investor's signal (τ_{ε_i}) is chosen at a cost described below. Since investors are homogeneous, one can show that traders choose the same precision level, i.e., $\tau_{\varepsilon_i} = \tau_\varepsilon \forall i$. This learning technology resembles Verrecchia (1982). Different from Grossman and Stiglitz (1980) or Hellwig (1980), in this setting, investors are *all* informed and decide the optimal precision of their signal. In other words, information acquisition acts on the intensive margin rather than the extensive margin. Finally, noise traders demand $\tilde{x} \sim N(0, \sigma_x^2)$, independent of all other random variables.

3.2.3 Data-Mining Technology

Investors are not endowed with informative datasets. They need to search for profitable signals (i.e., information not already impounded into stock prices) in the pool of all available information. The more the firm discloses, the more the total volume of data about the firm increases. However, the number of *profitable* signals does not necessarily rise one-to-one with public information. When public information increases faster than the stock of profitable signals, investors must exert more effort to find a profitable data point. This mechanism makes profitable data more valuable as public information grows.

Formally, the total volume of data available about the firm has intensity $N(\tau_{pub})$, with $N'(\tau_{pub}) > 0$: more public information generates more data about the firm. For example, once the firm discloses more information, that is, as τ_{pub} increases, the disclosure may attract media coverage or generate more discussion by commentators and analysts. This coverage increases the total amount of data available for processing.

Of this data, only a fixed stock $n_0 > 0$ consists of profitable signals that provide information about $\tilde{\theta}$ beyond the public signal ω . The remaining $N(\tau_{pub}) - n_0$ data points are redundant, their information content is already subsumed by ω or is incorporated in the price. As a result, the signal density $q(\tau_{pub}) := n_0/N(\tau_{pub})$ is decreasing. When public information rises, either because of more disclosure or more informed managers, the fraction of profitable data falls. This property is crucial for the mechanism in this paper. In Section 3.2.4, I discuss under what conditions this property is not satisfied and the implications for the main results.

Each investor processes raw data to construct her profitable signal. An investor who processes raw data must exert effort, which I denote as s_i . For example, s_i may represent the cost of hiring an analyst or the cost of training and backtesting a predictive model. By exerting effort s_i , the investor extracts $s_i \cdot q(\tau_{pub}) \cdot \tau_d$ units of precision from the raw data, where $\tau_d > 0$ is the precision of each profitable data point. The cost of processing raw data is $(c/2)s_i^2$, which is convex in the effort s_i . Substituting $s_i = \tau_\varepsilon/(q(\tau_{pub})\tau_d)$ into the cost function yields the effective cost of achieving total precision τ_ε :

$$\frac{k(\tau_{pub})}{2} \tau_\varepsilon^2, \quad \text{where} \quad k(\tau_{pub}) = k_0 \cdot r(\tau_{pub})^2, \quad (3.7)$$

with $k_0 := c/\tau_d^2$ and $r(\tau_{pub}) := N(\tau_{pub})/n_0 \geq 1$. The cost function $k(\tau_{pub})$ determines how costly it is to extract one piece of profitable information. Since $r'(\tau_{pub}) = N'(\tau_{pub})/n_0 > 0$, the effective cost $k(\tau_{pub})$ is increasing in public precision: $k'(\tau_{pub}) > 0$. More precise public information makes it more costly to find profitable signals in the pool of all available information.

The key parameter governing the comparative statics is the *cost semi-elasticity*

$$\eta := \frac{k'(\tau_{pub})}{k(\tau_{pub})} = \frac{2 N'(\tau_{pub})}{N(\tau_{pub})}, \quad (3.8)$$

which measures how fast public information inflates the cost to search for a profitable signal. When $N'(\tau_{pub}) = 0$ (public information does not affect total data availability about the firm), $k(\tau_{pub}) = k_0$ for all τ_{pub} , and the firm's disclosure does not affect the cost of generating profitable signals. In this case, the model reduces to the standard setting in [Verrecchia \(1982\)](#).

The cost function $k(\tau_{pub}) = k_0 r(\tau_{pub})^2$ can be seen as a fixed research cost k_0 and a variable premium in transforming raw data into profitable signals ($r(\tau_{pub})^2 - 1$). The second component captures how hard it is to locate useful signals. Consider for example a data vendor who has already identified the n_0 useful signals out of

the total $N(\tau_{pub})$, and sells the profitable signals with unit precision τ_d at a price p_d . Since an investor's alternative is to achieve the same precision paying the research cost $(k(\tau_{pub})/2) \tau_d^2$, the gains from trade between the vendor and the investor are

$$\frac{k(\tau_{pub}) - k_0}{2} \tau_d^2 = \frac{k_0(r(\tau_{pub})^2 - 1)}{2} \tau_d^2 = p_d, \quad (3.9)$$

which are increasing in τ_{pub} for a given signal quality τ_d . As public information increases, the value of buying informative signals rises. This is not because the data itself is more informative, but because the cost of searching through irrelevant data has increased. Therefore, the value of data may increase even absent an improvement in the quality of the underlying dataset.

3.2.4 Discussion of the Decline in Profitable Signals with τ_{pub}

The decline in the ratio of profitable signals with τ_{pub} is crucial for the main results of this paper. This decline can arise from several mechanisms that do not produce a one-to-one mapping between public information and the information from which investors can profit. For example, a firm's public disclosure may reveal information that was previously profitable for informed investors. Once the information is public, it is incorporated into the price and traders can no longer profit from it. Even if public disclosure did not affect the profitability of existing investors' signals at all, the fraction of profitable data would still fall with public information, and the paper's main mechanism would remain intact. The mechanism would fail, however, if public disclosure generated enough *new* profitable data to more than offset the information it reveals. For example, this might happen if markets violate weak market efficiency. If investors do not have the same ability to read the public signal, τ_{pub} might generate new investment opportunities from publicly available information.

Building on [Menzio \(2007\)](#), in Section 3.4.1, I discuss a similar case where investors have some skill in optimally directing their search towards more profitable data sources. I show that under this scenario, unless investors know how to perfectly filter profitable signals from unprofitable ones, the main results of the paper are unchanged.

Notably, this rising cost in information acquisition can also originate from limited attention. Suppose investors have a fixed capacity to process information, as in standard models of rational inattention ([Sims, 2003](#); [Peng and Xiong, 2006](#)). Every signal disclosed by the firm, together with the additional content it generates, competes for the scarce attention capacity. As τ_{pub} grows, the capacity left for private research shrinks and the marginal cost of achieving any research precision τ_e rises,

i.e., $k'(\tau_{pub}) > 0$. The underlying economics is the same as in this paper: public information absorbs a scarce resource that investors need in order to produce private information. Thus, the choice between the two settings would not change the paper's main results.

3.3 Equilibrium

The equilibrium is a standard NREE and consists of a price function and manager's optimal disclosure such that the market clears and all agents maximize expected utility conditional on their information sets. I solve for the symmetric linear equilibrium backwards: first, I solve for the NREE for a given common precision τ_ε , which implies the optimal investors' effort in searching for profitable data, then I find the manager's optimal disclosure choice. All proofs are in Appendix 3.6.1.

3.3.1 Market Equilibrium

Following the standard approach in the literature, I focus on a linear equilibrium. I conjecture and later verify a linear price function

$$\tilde{P} = p_0 + p_\omega \omega + p_\theta \tilde{\theta} + p_x \tilde{x}, \quad (3.10)$$

where the p -coefficients are endogenously determined in equilibrium. Since p_0 and ω are known to every investor, the price signal about $\tilde{\theta}$ is

$$z_P := \frac{P - p_0 - p_\omega \omega}{p_\theta} = \tilde{\theta} + \frac{p_x}{p_\theta} \tilde{x}, \quad (3.11)$$

with precision $\tau_z := p_\theta^2 / (p_x^2 \sigma_x^2)$. When τ_z is relatively high, the stock price carries precise information about the realization of the fundamental $\tilde{\theta}$. This can happen, for example, because the price is sensitive to changes in the fundamentals (p_θ is high), or because noise has little effect on the stock price (low p_x or σ_x^2). From the investors' perspective, observing the market price is informationally equivalent to observing the signal z_P . Therefore, an investor observes three mutually independent signals about $\tilde{\theta}$, namely ω , y_i , and z_P .

Investor i 's CARA-normal optimal demand is

$$D_i = \frac{\mathbb{E}[\tilde{\theta}|\omega, y_i, z_P] - \tilde{P}}{\rho \mathbb{V}ar[\tilde{\theta}|\omega, y_i, z_P]} = \frac{1}{\rho} \left[\tau_\varepsilon y_i + \tau_\omega \omega + \tau_z z_P - \Sigma \tilde{P} \right], \quad (3.12)$$

where I define the total market precision as $\Sigma := \tau_{pub} + \tau_\varepsilon + \tau_z$, which summarizes the total precision of market participants about $\tilde{\theta}$. Σ increases not only in the precision of public signals (τ_{pub} and τ_z), but also with the precision of investors' private information since they partially reveal their signals through the market price.

The market clearing condition $\int_0^1 D_i di + \tilde{x} = Q$ yields, after matching coefficients:

$$p_0 = -\frac{\rho Q}{\Sigma}, \quad p_\omega = \frac{\tau_\omega}{\Sigma}, \quad p_\theta = \frac{\tau_\varepsilon + \tau_z}{\Sigma}, \quad p_x = \frac{\rho + \tau_\varepsilon/(\rho \sigma_x^2)}{\Sigma}. \quad (3.13)$$

The coefficient p_0 is akin to the risk premium that market participants require to hold the firm's stock, on average. Investors require more compensation for holding the risky asset when they are more risk averse, when they are required to absorb a larger share of the firm's risk, or when they have relatively lower total market precision (Σ). The other endogenous coefficients p_ω , p_θ , and p_x represent the sensitivity of price with respect to variation in ω , $\tilde{\theta}$, and $\tilde{x}$, respectively. For example, as one would expect, the stock price is more sensitive to the manager's public disclosure when the disclosure itself is more informative.

Substituting the coefficients in Equations (3.13) into $\tau_z = p_\theta^2/(p_x^2 \sigma_x^2)$ gives

$$\tau_z = \frac{\tau_\varepsilon^2}{\rho^2 \sigma_x^2}, \quad (3.14)$$

which expresses the precision of the price signal as a function of the primitive parameters. The accuracy of the price signal, τ_z , depends on the precision of investors' private information (τ_ε), on how aggressively they trade on it (through ρ), and on the noise σ_x . Crucially, even though the precision of the price signal does not depend directly on the manager's public disclosure, τ_z is linked indirectly to τ_v (and τ_{pub}) through the investors' equilibrium precision of the private signal.

In fact, each investor chooses τ_ε to maximize her certainty-equivalent gain minus the (endogenous) cost of searching for profitable signals, taking the price-signal

precision τ_z as given:

$$\max_{\tau_\varepsilon \geq 0} \frac{1}{2\rho} \ln \left(\frac{\tau_{pub} + \tau_\varepsilon + \tau_z}{\tau_{pub} + \tau_z} \right) - \underbrace{\frac{k(\tau_{pub})}{2} \tau_\varepsilon^2}_{\text{search cost}}. \quad (3.15)$$

The first-order condition in the symmetric equilibrium ($\tau_{\varepsilon,i} = \tau_\varepsilon$ for all i) is

$$\frac{1}{2\rho \Sigma} = k(\tau_{pub}) \cdot \tau_\varepsilon, \quad \text{i.e.,} \quad \tau_\varepsilon^* = \frac{1}{2\rho k(\tau_{pub}) \Sigma}. \quad (3.16)$$

In the linear equilibrium, this solution for τ_ε^* always exists and is unique. The absence of a fixed cost ensures that every investor acquires at least some information, in contrast to the extensive-margin formulation (e.g., [Grossman and Stiglitz, 1980](#)) where a fraction of investors remain uninformed. Moreover, investors exert less effort in acquiring private information when there is more public information available. This is the standard crowding-out of public information of [Verrecchia \(1982\)](#) and [Diamond \(1985\)](#). See also [Goldstein and Yang \(2017a\)](#) for a review. Since public information gets impounded into the market price, investors have lower incentives to acquire private information. In equilibrium, the decrease in the private signal's precision exactly offsets the increase in public information.

However, in this paper, there is another economic mechanism that crowds out private information acquisition. When the amount of public signals increases more than the pool of truly *profitable* data, investors find it even more costly to search for useful signals. In the words of [Dugast and Foucault \(2025\)](#), when firms produce more public information (either directly through disclosure or indirectly through media coverage) it becomes harder for investors to find the needle in the haystack. Indeed, denoting for simplicity $\Gamma := \tau_{pub} + 2\tau_\varepsilon + 3\tau_z$, the following three results always hold:

$$\frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} = -\frac{\tau_\varepsilon}{\Gamma} (1 + \eta \Sigma) < 0, \quad (3.17)$$

$$\frac{\partial \tau_z}{\partial \tau_{pub}} = -\frac{2\tau_z}{\Gamma} (1 + \eta \Sigma) < 0, \quad (3.18)$$

$$\frac{\partial \Sigma}{\partial \tau_{pub}} = \frac{\Sigma [1 - (\tau_\varepsilon + 2\tau_z) \eta]}{\Gamma}. \quad (3.19)$$

Appendix 3.6.1 reports the derivations. Equation (3.17) shows that the investors' research precision decreases when public precision rises. This result is similar to previous results on the crowding-out of private information ([Verrecchia, 1982](#); [Diamond, 1985](#)). Yet, when the ratio of profitable signals to the total amount of

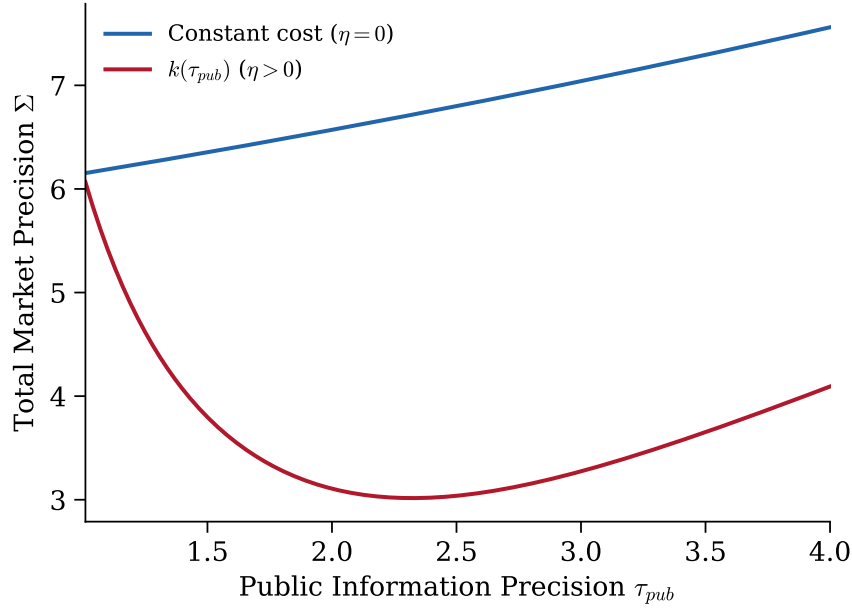


FIGURE 3.2: **Total Market Precision and Public Information.** This figure shows the total market precision $\Sigma = \tau_{pub} + \tau_\varepsilon + \tau_z$ as a function of the precision of public information τ_{pub} . The blue line represents the benchmark with constant cost of information acquisition ($\eta = 0$): Σ is monotonically increasing, so the gain in public precision always exceeds the crowding-out of private research. The red line represents the endogenous search cost $k(\tau_{pub})$ proposed in this paper ($\eta > 0$): Σ initially decreases, as investors find it more costly to search for profitable signals within the pool of all available data. The parameters for this Figure are: $k_0 = 0.08$, $\gamma = 1.2$, $\sigma_x = 0.5$, $\rho = 1$, $\tau_\theta = 1$.

information also decreases with public information, the precision of the price signal declines as well (Equation (3.18)). The term in parentheses, $(1 + \eta\Sigma)$, exceeds one whenever $\eta > 0$: as investors find it more difficult to search for profitable data, the crowding-out of private research is even stronger. In Section 3.3.3, I study the implication of this mechanism for price informativeness.

The sign of $\partial\Sigma/\partial\tau_{pub}$ depends on η . When $\eta = 0$, as in the case of constant cost, which is the standard choice in the literature, $\partial\Sigma/\partial\tau_{pub} = \Sigma/\Gamma > 0$: the total market precision is always increasing in public information. However, when the cost of acquiring information varies with the amount of public information ($\eta > 0$) the total precision of market participants can *decrease* in τ_{pub} . This marginal decline in total market information arises when the cost semi-elasticity η is large enough—in particular, when $(\tau_\varepsilon + 2\tau_z)\eta > 1$. In this case, filtering useful signals from the pool of available data is so costly that public information crowds out private research more than one for one. Figure 3.2 illustrates the two cases. Crucially, the combined loss in τ_ε and τ_z can exceed the gain in τ_{pub} . When disclosure generates enough derivative data, public information creates a negative externality for the information produced by the economy. The total amount of data inflates faster than the increase in profitable data sources, and investors' optimal research effort shrinks.

3.3.2 Optimal Disclosure

The market equilibrium so far has taken the disclosure precision τ_v as given. I now solve backward for the manager's optimal disclosure, taking into account how her choice feeds back into investors' research precision and the informational content of the stock price.

In $t = 0$, the manager chooses the precision of disclosure τ_v to maximize the expected stock price net of the proprietary cost of disclosure

$$\max_{\tau_v} \mathbb{E}[\tilde{P}] - C(\tau_v). \quad (3.20)$$

Since $\mathbb{E}[\omega] = \mathbb{E}[\tilde{\theta}] = \mathbb{E}[\tilde{x}] = 0$, the expected stock price reduces to

$$\mathbb{E}[\tilde{P}] = p_0 = -\frac{\rho Q}{\Sigma}. \quad (3.21)$$

Higher total market precision Σ reduces the risk premium and raises the expected stock price. The manager's problem is therefore to choose τ_v to maximize $-\rho Q/\Sigma - C(\tau_v)$, where Σ depends on τ_v . The marginal benefit of an increase in disclosure precision is given by

$$\frac{\partial p_0}{\partial \tau_v} = \frac{\rho Q}{\Sigma^2} \frac{\partial \Sigma}{\partial \tau_{pub}} \frac{\partial \tau_{pub}}{\partial \tau_v}. \quad (3.22)$$

Since $\partial \tau_{pub}/\partial \tau_v = \tau_\chi^2/(\tau_\chi + \tau_v)^2 > 0$, the sign of this marginal benefit is the sign of $\partial \Sigma/\partial \tau_{pub}$, which is given in Equation (3.19). When $\partial \Sigma/\partial \tau_{pub} > 0$, from the manager's perspective, public disclosure raises the expected stock price. In this case, the reduction in risk premium exceeds the loss from the crowding-out of private information. Therefore, the manager finds it marginally optimal to publicly disclose part of her internal information. On the other hand, when $\partial \Sigma/\partial \tau_{pub} < 0$, disclosure *lowers* the expected stock price as the total amount of data available to investors for processing increases so much that the total market precision falls. In this case, the manager never prefers to crowd out the private information generated by investors' search for profitable signals and will not disclose part of her internal information.

Proposition 3.1 (Optimal disclosure). *The manager's optimal disclosure problem (Equation 3.20) always has a unique interior optimum $\tau_v^* > 0$, as long as $\partial \Sigma/\partial \tau_{pub} > 0$ at the starting point $\tau_{pub} = \tau_\theta$ (i.e., when $\tau_v = 0$). The optimal disclosure solution satisfies $\partial p_0/\partial \tau_v^* = C'(\tau_v^*)$.*

The proof is in Appendix 3.6.1. When the manager sets no public disclosure ($\tau_v = 0$), $C'(0) = 0$ and $\partial p_0/\partial \tau_v|_{\tau_v=0} > 0$ (since $\partial \Sigma/\partial \tau_{pub} > 0$ at $\tau_{pub} = \tau_\theta$), so the

manager wants to marginally increase her disclosure from zero. As the manager increases the precision of her public disclosure, the marginal benefit deteriorates, since $\partial\tau_{pub}/\partial\tau_v \rightarrow 0$ as τ_v grows, while the marginal disclosure cost $C'(\tau_v)$ increases. Therefore, there is a unique optimal disclosure τ_v^* that maximizes the expected stock price (or minimizes the firm's expected cost of capital).

The condition $\partial\Sigma/\partial\tau_{pub} > 0$ at $\tau_{pub} = \tau_\theta$ ensures that public disclosure raises the total market precision at zero disclosure, and the manager finds it optimal to disclose at least some information. This condition holds when $(\tau_\varepsilon + 2\tau_z)\eta < 1$, at $\tau_{pub} = \tau_\theta$. That is, when the ratio between public and profitable signals increases moderately at low levels of public precision. This is a plausible assumption for firms that have not yet disclosed any information. For example, media coverage plausibly “inflates” total public information more for firms with higher ex-ante disclosure than for opaque firms.

Figure 3.3 shows the manager's payoff as a function of her disclosure precision τ_v , for different levels of firm's internal information τ_χ . The optimal disclosure shifts rightward as the firm has more precise internal information, i.e., τ_χ increases. When the firm holds more internal data unavailable to the market, the manager optimally discloses part of what she learns from it. Even though this public disclosure possibly crowds out investors' effort in searching for useful information about the firm's fundamentals, the manager benefits from lower (expected) cost of capital.

Notably, the manager has an interior optimal disclosure even without a direct disclosure cost $C(\cdot)$. Beyond a threshold, public information grows disproportionately relative to the stock of profitable data, causing investors to exert no effort in searching for mispricing. In this case, the total market precision deteriorates and the firm's cost of capital rises (the expected stock price declines). In traditional models with constant cost of acquiring information (e.g., Grossman and Stiglitz, 1980; Verrecchia, 1982), the manager either discloses *all* the information learned from internal data or nothing. However, the previous result shows that if investors face increasing cost of searching for profitable signals when public information rises, managers partially disclose the information learned from data. This result implies that firms' current investments in new technologies that improve their internal knowledge spill over to the entire market via endogenous disclosure. Frésard and Zanotti (2026) find evidence consistent with this result. Using I/B/E/S Guidance data on managers' voluntary disclosure, they show that once US retail firms adopt technologies aimed at collecting and analyzing customers' data, they are about 7% more likely to issue forward guidance on future earnings and sales. Moreover, for firms that were already reporting before the adoption of data analytics technologies, the voluntary disclosure

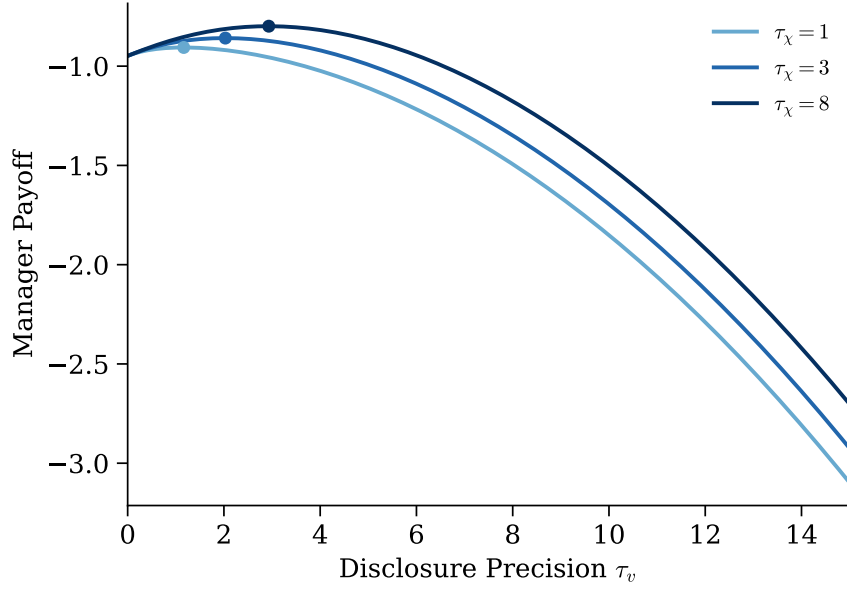


FIGURE 3.3: **The Manager's Optimal Disclosure.** This figure shows the manager's payoff as a function of the precision of public disclosure τ_v , for three levels of the firm's internal data precision $\tau_\chi \in \{1, 3, 8\}$. Darker colors represent higher levels of internal data precision τ_χ . The dots indicate the manager's optimal disclosure τ_v^* for each level of τ_χ . More internal data shifts the optimal disclosure to the right. The parameters for this Figure are: $k_0 = 0.15$, $\gamma = 0.05$, $\sigma_x = 1$, $Q = 3$, cost of disclosure $c_d = 0.02$, $\rho = 1$, $\tau_\theta = 1$.

becomes more precise. This is consistent with managers raising the precision of public disclosure τ_v as a response to more data availability.

Interestingly, the manager discloses even more precise information when the market is noisier, i.e., when σ_x is high. Formally, $\partial \tau_v^* / \partial \sigma_x > 0$. When σ_x is high, the precision of the price signal $\tau_z = \tau_\epsilon^2 / (\rho^2 \sigma_x^2)$ is relatively small as investors are not able to disentangle fundamentals from noise. The total precision of market participants Σ is low, and thus the required risk premium $\rho Q / \Sigma$ is high. In this scenario, each unit of disclosure delivers a relatively larger reduction in the firm's cost of capital. As a result, when σ_x is high, managers find it even more valuable to communicate with the public and disclose part of their internal information. Figure 3.4 depicts this result. Optimal disclosure rises more steeply with the firm's internal information τ_χ in noisier markets (higher σ_x).

3.3.3 Price Informativeness

Having characterized the manager's optimal disclosure, I now turn to what the firm's data implies for the informativeness of the stock price. I define price informativeness as $\Phi := \tau_{pub} + \tau_z$. This definition is equivalent to Forecasting Price Efficiency (FPE) in [Bond et al. \(2012\)](#). Φ captures the total precision available to an outside observer who conditions her belief about the fundamental $\tilde{\theta}$ on the price

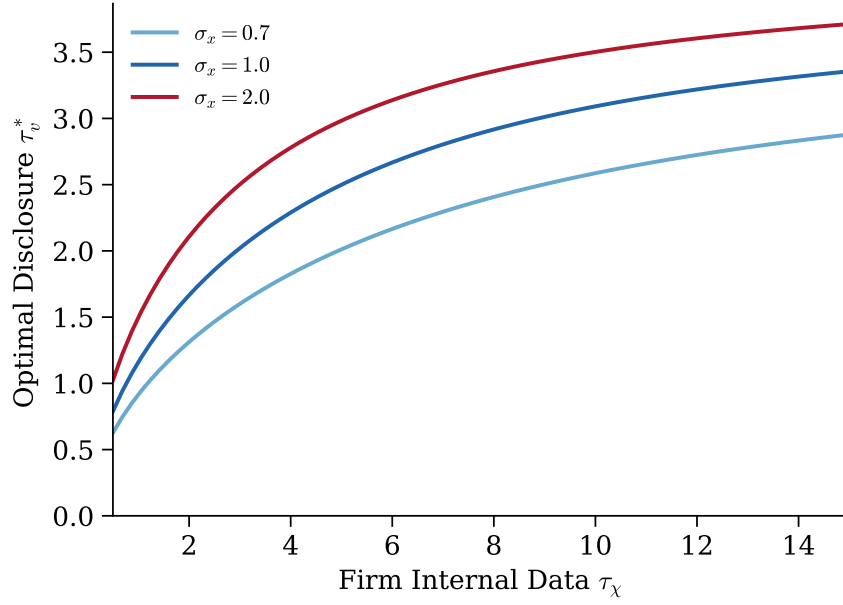


FIGURE 3.4: **Optimal Disclosure and Stock Market Noise.** This figure shows the optimal disclosure precision τ_v^* as a function of the firm's internal information τ_χ , for three levels of stock market noise $\sigma_x \in \{0.7, 1.0, 2.0\}$ (light blue, dark blue, and red lines, respectively). Optimal disclosure increases with the firm's internal information, and its slope is steeper in markets with more noise trading (higher σ_x), where each unit of disclosure delivers a larger reduction in the firm's cost of capital. The parameters for this Figure are: $k_0 = 0.15$, $\gamma = 0.05$, $Q = 3$, $c_d = 0.02$, $\rho = 1$, $\tau_\theta = 1$.

and public information. In standard models with an intensive margin of information acquisition and constant cost, Φ is always increasing in τ_{pub} . As in [Verrecchia \(1982\)](#), the gain in public precision exceeds the crowding-out loss in τ_z , and more public information unambiguously makes the price more informative. This result does not hold when the cost of acquiring information endogenously increases with the firm's public disclosure of informative signals.

Proposition 3.2 (Price informativeness). *In the symmetric equilibrium:*

$$\frac{\partial \Phi}{\partial \tau_{pub}} = \frac{\tau_{pub} + 2\tau_\varepsilon + \tau_z - 2\tau_z \eta \Sigma}{\Gamma}. \quad (3.23)$$

Price informativeness is decreasing in the precision of the firm's internal data if and only if

$$\tau_z > \frac{\tau_{pub} + 2\tau_\varepsilon}{2\eta \Sigma - 1}. \quad (3.24)$$

When $\eta = 0$ (constant cost), $\partial \Phi / \partial \tau_{pub} = (\tau_{pub} + 2\tau_\varepsilon + \tau_z) / \Gamma > 0$: FPE is always increasing.

The condition (3.24) identifies when the increase in search cost dominates. This

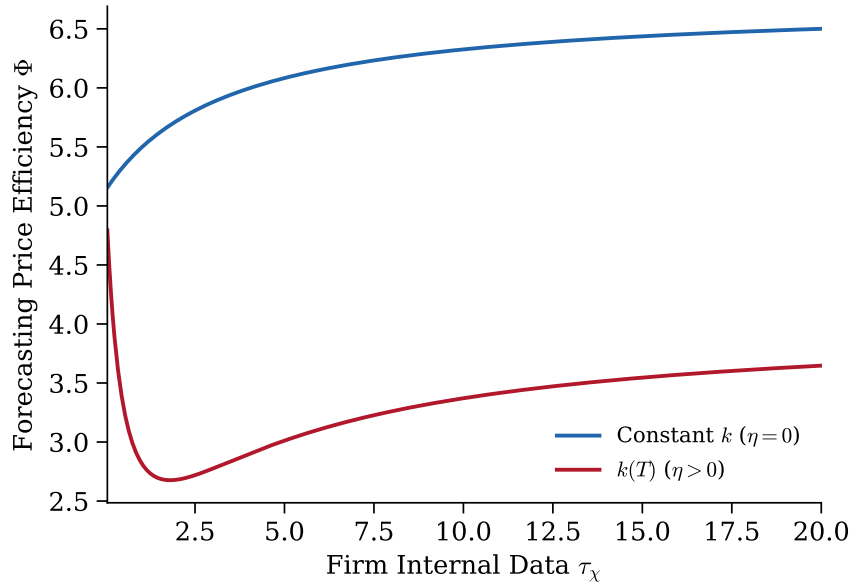


FIGURE 3.5: **Price Informativeness and the Firm's Internal Data.** This figure shows forecasting price efficiency $\Phi = \tau_{pub} + \tau_z$ as a function of the firm's internal information τ_χ . The blue line represents standard models with constant cost of information acquisition ($\eta = 0$) where Φ is always monotonically increasing in the firm's internal data. The red line represents the endogenous search cost proposed in this paper ($\eta > 0$). The parameters for this Figure are: $k_0 = 0.08$, $\gamma = 1.2$, $\sigma_x = 0.5$, $\tau_v = 3$, $\rho = 1$, $\tau_\theta = 1$.

condition can be rewritten as

$$\tau_{pub} + 2\tau_\varepsilon + \tau_z - 2\tau_z \eta \Sigma > 0.$$

The $-2\tau_z \eta \Sigma$ term measures the cost of a reduction in τ_z . This component is large when η is large (i.e., the search cost increases rapidly with total information), or when τ_z is large to begin with. The rest of the equation captures the direct effects on price informativeness of an increase in the precision of public information and the standard crowding-out response. When the cost term exceeds these effects, the total information available in the market's price deteriorates. I show in Figure 3.5 how price informativeness varies with firms' internal data. The blue line represents the standard case with constant cost of acquiring information, while the red line shows this paper's main result. The possibility that the total information contained in the price (FPE) decreases with public information (and public disclosure) is unique to this setting and contrasts sharply with the predictions of standard models of endogenous information acquisition and disclosure.

The externality of public information changes every equilibrium object, not just FPE. Figure 3.6 compares the full equilibrium under the endogenous search cost ($\eta > 0$) with the constant-cost benchmark ($\eta = 0$). Panels (a) to (c) trace the mechanism: as the firm's internal data τ_χ increases, the premium for finding profitable data points rises,

the search cost rises with it, and the crowding-out of private research is substantially stronger than under constant cost. Panels (d) to (f) show the market consequences: price informativeness initially decreases, the risk premium is non-monotone in the firm's internal information, and total market precision grows more slowly than under constant cost because the gain in public precision is partially offset by the larger loss in private information.

The possibility that price informativeness declines with the firm's internal data yields a testable implication. Once equipped with good measures of the extent to which the market price reflects future fundamentals, one could test whether stock price informativeness deteriorates once firms have access to more precise internal data. This prediction is in stark contrast with the common argument in the literature that more data unambiguously improves price informativeness. For example, ? and ? find that investors' access to more data in the past decades has increased FPE, especially for large and growth firms. The important difference, however, is that in this paper I study what happens to price informativeness when *firms*, rather than investors, have access to more data. I show that the impact of data on price informativeness is not always positive, as it is crucial to recognize that firms, too, can access more precise sources of data.

3.3.4 Cross-Firm Externalities

I now extend the model to a two-firm economy, to study how data adoption at one firm can affect information production at another. This extension generates testable predictions for how data adoption can indirectly affect product market peers and their stock prices.

Similar to [Tookes \(2008\)](#), there are now two firms $i \in \{A, B\}$ with independent fundamentals $\tilde{\theta}_i \sim N(0, 1/\tau_\theta)$. Each firm has the same structure as above: the manager observes a signal s_i with precision $\tau_{\chi,i}$ and discloses with precision $\tau_{v,i}$. For simplicity, both firms have the same total supply of risky shares Q . Each firm's disclosure generates a total amount of public data $N(\tau_{pub,i})$, with the ratio of public over profitable signals given by $r_i := r(\tau_{pub,i}) = N(\tau_{pub,i})/n_0$.

Investors acquire two private signals, one about each firm $(\tau_{\varepsilon,A}, \tau_{\varepsilon,B})$. Similarly to before, the total cost of searching for profitable information in the pool of raw data is

$$\frac{k_0}{2} (r_A \tau_{\varepsilon,A} + r_B \tau_{\varepsilon,B})^2, \quad (3.25)$$

where $r_i \tau_{\varepsilon,i}$ is the search effort required to achieve precision $\tau_{\varepsilon,i}$ on firm i . The quadratic cost implicitly captures gains from specialization and limited research

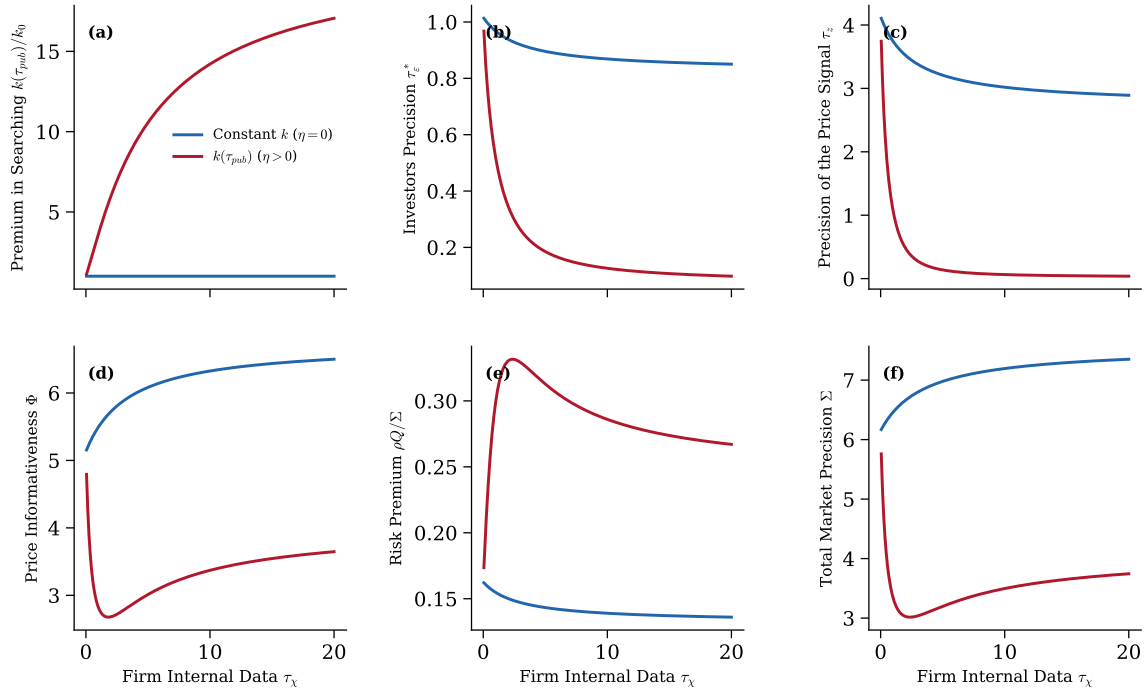


FIGURE 3.6: **Equilibrium Comparative Statics.** This figure shows comparative statics of the equilibrium as a function of the firm's internal data τ_χ , comparing standard models with constant cost of information acquisition ($\eta = 0$, blue lines) with the endogenous search cost proposed in this paper ($\eta > 0$, red lines). Panel (a) depicts the premium in searching for profitable signals $k(\tau_{pub})/k_0$. Panels (b) and (c) show the equilibrium research precision τ_ε^* and the precision of the price signal τ_z . Panel (d) shows price informativeness $\Phi = \tau_{pub} + \tau_z$, Panel (e) the risk premium $\rho Q/\Sigma$, and Panel (f) the total market precision Σ . The parameters for this Figure are: $k_0 = 0.08$, $\gamma = 1.2$, $\sigma_\chi = 0.5$, $\tau_v = 3$, $Q = 1$, $\rho = 1$, $\tau_\theta = 1$.

capacity, as devoting more effort to one firm raises the marginal cost of researching the other.

Since $\tilde{\theta}_1 \perp \tilde{\theta}_2$, each firm's market equilibrium is a separate NREE with $\tau_{z,i} = \tau_{\varepsilon,i}^2/(\rho^2\sigma_\chi^2)$ and $\Sigma_i = \tau_{pub,i} + \tau_{\varepsilon,i} + \tau_{z,i}$. However, in this two-firm economy, the investor's first-order conditions for the two private signals' precisions imply an equilibrium condition:

$$r_A \cdot \Sigma_A = r_B \cdot \Sigma_B. \quad (3.26)$$

In equilibrium, the cost-adjusted total market precision across the two firms must be the same. In the symmetric case where both firms have the same internal information and public disclosure choice ($\tau_{pub,A} = \tau_{pub,B}$, hence $r_A = r_B$), this condition reduces to $\Sigma_A = \Sigma_B$.

Proposition 3.3 (Cross-firm spillover). *Starting from a symmetric equilibrium, suppose firm A adopts a new technology that increases the precision of its internal information ($\tau_{\chi,A}$) while firm B does not adopt. Then, in equilibrium, investors exert less effort to acquire private information on firm A ($\partial\tau_{\varepsilon,A}/\partial\tau_{\chi,A} < 0$) and reallocate that effort to generate information*

about firm B ($\partial\tau_{\varepsilon,B}/\partial\tau_{\chi,A} > 0$). As a result, firm B's price informativeness improves ($\partial\Phi_B/\partial\tau_{\chi,A} > 0$) and firm B's cost of capital decreases ($\partial\mathbb{E}[\tilde{P}_B]/\partial\tau_{\chi,A} > 0$).

This mechanism operates through two channels. The first is the standard channel described above: as firm A's internal information raises $\tau_{pub,A}$, investors' marginal value of searching for profitable signals about firm A falls. Second, higher public information on firm A increases the effective cost of research on firm A *relative* to firm B. Since investors have a fixed amount of effort to allocate between the two firms, they find it optimal to reallocate effort toward firm B. Because r_A has increased, condition (3.26) requires Σ_A to fall and Σ_B to rise.

3.4 Extension and Discussion

The results so far rest on a reduced-form assumption: the cost of acquiring private information rises with the volume of public data around the firm (Equation (3.7)). This section microfoundations that assumption with a sequential search model, shows that the externality survives—although attenuated—when investors can partially direct their search, and discusses whether its magnitude is empirically plausible.

3.4.1 Microfoundation: Sequential Search of Profitable Data

The endogenous cost of acquiring information is determined by $k(\tau_{pub}) = k_0 \cdot r(\tau_{pub})^2$, see Equation (3.7). In Section 3.2, this cost structure rests on the assumption that the density of profitable signals over total public information $q(\tau_{pub}) = n_0/N(\tau_{pub})$ falls as public precision rises. I now provide a microfoundation for this structure using a sequential search model inspired by the literature on search costs in markets with heterogeneous goods (Menzio, 2007; Menzio and Trachter, 2015).

As before, the total number of public data points around the firm consists of $N(\tau_{pub})$ items. Each data point is either informative, or uninformative (e.g., redundant). *Informative data points* (n_0 in number) have unit precision $\tau_d > 0$ about the fundamental $\tilde{\theta}$, beyond what public information ω already reveals. These data points exist independently of the firm's disclosure and they represent proprietary information generated by investors. For example, these might represent satellite data, supply-chain information, or other sources whose information content allows the investor to profit. *Uninformative data points* ($N(\tau_{pub}) - n_0$), on the other hand, carry zero additional precision about $\tilde{\theta}$ beyond ω . These items are derivative of the firm's public information. For instance, they represent commentary on the firm's

disclosure, sell-side summaries, or media articles. In general, these data points restate, interpret, or repackage publicly available information so they do not contain additional information beyond what is contained in ω . Crucially, an investor cannot distinguish the two types a priori without inspection of the raw data.

Content Creation and Free Entry. The number of uninformative (redundant) data points is endogenous. There is a competitive content creation sector (e.g., the media sector) that takes the firm's disclosure ω as a free input and produces derivative data at cost $f > 0$ per item. The cost of producing a data point represents the infrastructure or human capital cost of interpreting and repackaging the disclosed signal. Each data point generates revenue from investors' attention. A single raw data point inspected by investors generates unit revenue $\pi(N, \tau_{pub})$, where π is decreasing in N (attention per item falls as the pool of raw data points grows) and increasing in τ_{pub} (richer disclosures support the business of creating more differentiated content). The business of creating content has free entry cost. Thus, in equilibrium

$$\pi(N(\tau_{pub}), \tau_{pub}) = f. \quad (3.27)$$

Since $\partial\pi/\partial\tau_{pub} > 0$ and $\partial\pi/\partial N < 0$, it holds that $N'(\tau_{pub}) = -\partial\pi/\partial\tau_{pub}/\partial\pi/\partial N > 0$. In other words, more public precision attracts more content creation, expanding the total amount of data points generated about the firm. The total stock of informative data points n_0 , however, does not respond to τ_{pub} because content creators do not generate profitable signals, but only content derived from public information. Therefore, the firm's disclosure (or internal information) cannot create new private information for investors.

Search Cost of Profitable Data. The investor searches for informative data points by sampling *sequentially* from raw data, i.e., searching for useful signals through $N(\tau_{pub})$ data points. Each search has a uniform probability $q(\tau_{pub}) = n_0/N(\tau_{pub})$ to draw a profitable signal, and with probability $1 - q(\tau_{pub})$ the data point drawn is already known from public information. The investor inspects or backtests each data point and, if it is informative, incorporates its precision τ_d into her private signal.

There is a cost to process raw data. As in the main model in Section 3.2, the total cost of processing S data points is $(c/2)S^2$. To reach the total signal precision τ_ε , the investor needs to find $m = \tau_\varepsilon/\tau_d$ informative data points³. Since each draw yields an

³Investors are rational and update their prior according to Bayes' law. Since private signals are independent, the total precision of m profitable data points is $m\tau_d$.

informative item with probability $q(\tau_{pub})$, the average number of draws is

$$S = \frac{\tau_\varepsilon}{\tau_d q(\tau_{pub})} = \frac{\tau_\varepsilon r(\tau_{pub})}{\tau_d}. \quad (3.28)$$

The total cost of achieving precision τ_ε is therefore

$$\frac{c}{2} S^2 = \frac{c r(\tau_{pub})^2}{2 \tau_d^2} \tau_\varepsilon^2 = \frac{k(\tau_{pub})}{2} \tau_\varepsilon^2, \quad (3.29)$$

where $k(\tau_{pub}) = k_0 \cdot r(\tau_{pub})^2$ with $k_0 := c/\tau_d^2$, delivering exactly the cost function as in Equation (3.7). The derivation is in Appendix 3.6.2.

In this simple setting, there are two important features. First, redundant data points are endogenous. The free entry condition in the business of creating public content generates more signals about the firm. Even if, from an informational point of view, these signals are pure noise, investors need to pre-process and inspect them to realize they do not contain additional information. Second, the search cost grows with $r(\tau_{pub})^2$ because public information creates an externality. Each extra public data point reduces the probability that a given draw is informative. This simple externality forces the investor to search longer at increasing marginal cost.

Partially Directed Search. The baseline derivation assumes that the investor draws uniformly a data set from the pool of all available signals, i.e., random search. In practice, investors might have some ability to understand which datasets are more likely to contain useful information. For example, sophisticated investors may realize that satellite data could be more valuable for searching investment opportunities than analysts guidance—which could partially represent a byproduct of publicly available information. Therefore, it is plausible that investors know how to *direct* their search toward more promising data sources. The partially directed search framework of Menzio (2007) provides a natural formalization of this.

The investor has access to a screening technology with accuracy $\delta \in [1/2, 1]$. The screening technology determines the ability of the investor to direct the search to promising data sources. Before inspecting a data point, she observes a binary signal: an informative data point is flagged “likely useful” with probability δ , while a redundant item is incorrectly flagged “likely useful” with probability $1 - \delta$. When $\delta = 1/2$, the screening technology is uninformative and the framework converges to the random search case analyzed above. On the other hand, when $\delta = 1$, screening is perfect and there are no externalities of public information, since the investor restricts her search only to data flagged “likely useful.”

The entire raw data contains $(n_0 \delta)$ true positives (correctly flagged as informative

signals) and $(N(\tau_{pub}) - n_0)(1 - \delta)$ false positives (uninformative signals that nonetheless passed the screening). Thus, as the investor processes only data points flagged as likely informative, the effective signal density within the screened pool of raw data is

$$q_{\text{eff}}(\tau_{pub}, \delta) = \frac{n_0 \delta}{n_0 \delta + (N(\tau_{pub}) - n_0)(1 - \delta)}, \quad (3.30)$$

and the effective ratio of processed data to profitable data becomes

$$r_{\text{eff}}(\tau_{pub}, \delta) = \frac{1}{q_{\text{eff}}} = 1 + (r(\tau_{pub}) - 1) \cdot \frac{1 - \delta}{\delta}. \quad (3.31)$$

Notice that when $\delta = 1/2$, Equation (3.31) reduces to the baseline specification above $r_{\text{eff}} = r(\tau_{pub})$. At the other extreme, when $\delta = 1$, the investor only processes useful datasets, i.e., $r_{\text{eff}} = 1$. Replacing $r(\tau_{pub})$ with $r_{\text{eff}}(\tau_{pub}, \delta)$ in the baseline search cost yields the generalized cost function when investors can partially direct their search effort:

$$k(\tau_{pub}, \delta) = k_0 \cdot r_{\text{eff}}(\tau_{pub}, \delta)^2. \quad (3.32)$$

The economic intuition is the same as in the baseline model in Section 3.2. Here, however, investors' screening technology partially mutes the externality of public information. As long as investors need to inspect a dataset to understand its quality (i.e., as long as $\delta < 1$), all the results of the paper hold. For example, the condition under which price informativeness declines with the firm's private information (Equation 3.24) holds for sufficiently large $N(\tau_{pub})$. Partially directed search attenuates the externality, but does not completely solve it as long as $\delta < 1$. The absolute number of false positives, $(N(\tau_{pub}) - n_0)(1 - \delta)$, grows with $N(\tau_{pub})$ even when the false-positive rate $1 - \delta$ is small. A screening technology with 90% accuracy ($\delta = 0.9$) still lets through 10% of the $N(\tau_{pub}) - n_0$ uninformative data points, and as the total amount of data points grows, these false positives eventually overwhelm the fixed stock of true positives in the screened pool.

This extension addresses the natural objection that investors do not search randomly across data vendors. Quantitatively skilled investors are likely to direct their search toward high-quality data sources, or they can use recent tools such as NLP or AI to pre-filter data. This result shows that such technologies raise the threshold of public information at which the decline of price informativeness occurs, but as long as the screening technologies are not perfect, they cannot prevent it.

Figure 3.7 shows how partially directed search attenuates the main results, for different levels of the screening technology δ . Panel (a) plots the cost semi-elasticity $\eta(\delta)/\eta(1/2)$ as a function of the screening accuracy δ . The semi-elasticity η captures

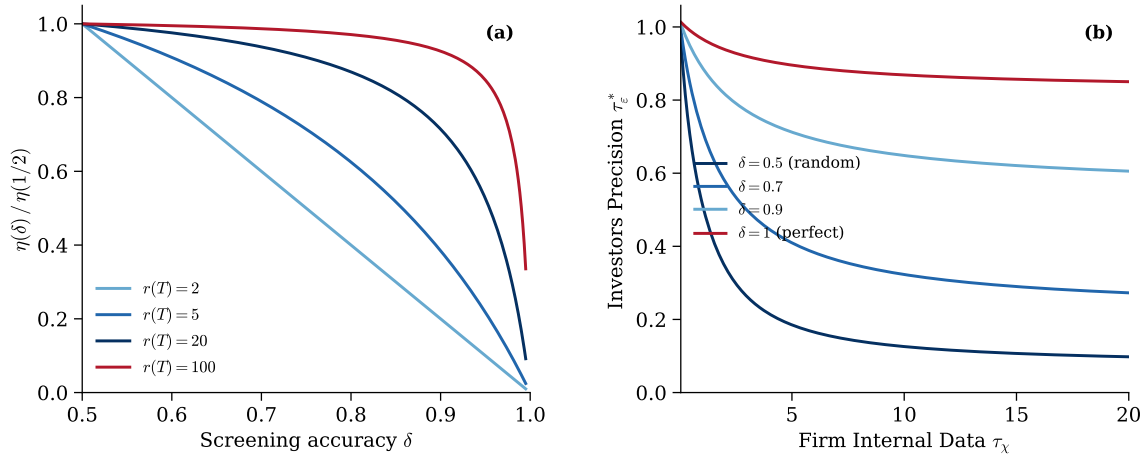


FIGURE 3.7: **Partially Directed Search.** This figure shows how screening technologies partially attenuate the externality of public information. Panel (a) plots the cost semi-elasticity $\eta(\delta)$, normalized by the baseline case with random search $\eta(1/2)$, as a function of the screening accuracy δ , for different volumes of the data available about the firm $r(\tau_{pub}) \in \{2, 5, 20, 100\}$. For environments with large volumes of data ($r(\tau_{pub}) = 100$, red line), $\eta(\delta)$ remains close to the random search case for all $\delta < 1$. Panel (b) reports the equilibrium research precision τ_ϵ^* as a function of the firm's internal data τ_χ , for different screening technologies $\delta \in \{0.5, 0.7, 0.9, 1\}$. The $\delta = 1$ curve (perfect screening, red line) coincides with traditional models with constant cost, while lower values of δ produce a stronger externality of public information. The parameters for this Figure are: $k_0 = 0.08$, $\gamma = 1.2$, $\sigma_x = 0.5$, $\tau_v = 3$, $\rho = 1$, $\tau_\theta = 1$.

how fast the search cost increases with more public information. When there are a few public signals available ($r(\tau_{pub}) = 2$), even moderate screening technology substantially reduces η . However, for firms with large volumes of public data available ($r(\tau_{pub}) = 100$), the ratio stays close to 1 across all $\delta < 1$. In other words, the semi-elasticity η converges to the random search case. Panel (b) shows the equilibrium precision for different levels of δ . As one might expect, private signal's precision τ_ϵ^* increases as a function of the quality of the screening technology. The $\delta = 1$ curve (perfect screening) coincides with the constant cost benchmark in traditional models, while lower values of δ produce a stronger externality of public information.

The partially directed search in this section might also generate cross-sectional predictions. For example, if the access to (or skills in using) the screening technologies δ varies across investors, quantitative investors might cope better with the proliferation of public data sources than traditional fundamental investors. The disclosure of public information could disproportionately harm investors with weaker screening ability, because their effective $r(\cdot)$ grows faster.

3.4.2 Discussion and Implications

The main distinctive element that summarizes this paper's main mechanism is the cost semi-elasticity $\eta = k'(\tau_{pub})/k(\tau_{pub})$, which summarizes the speed at which public

information raises the search cost of finding profitable signals. A natural question is whether η is empirically large enough for the externality of public information to be economically important.

Although a precise estimate of η is hard to obtain in practice—we typically do not observe investors' information sets—recent evidence suggests that η might be large. [de Haan et al. \(2025\)](#) construct an AI-based analyst that processes only publicly available data—e.g., filings, news, transcripts, and structured databases—to generate stock recommendations. They estimate that the marginal information-processing cost faced by this analyst is approximately \$17.1 million per quarter per fund. This cost exceeds both the average fund's alpha multiplied by its Assets Under Management (about \$2.8 million per quarter) and its management fees (about \$3.6 million per quarter) by a factor of five to six. In this model's language, this figure is a monetary proxy for $(k(\tau_{pub})/2) \tau_d^2$ (see Equation 3.9) for a given precision τ_d . It represents the cost of extracting profitable information from publicly available signals. The fact that this cost significantly affects the return to private information suggests that the premium for finding profitable data $r(\tau_{pub})^2 - 1$ is large. The total amount of data available on a single firm has expanded far beyond the stock of genuinely useful signals. Viewed differently, this number suggests that the market is extremely accurate in impounding publicly available information into the stock price.

These magnitudes can be connected to the condition for the negative impact of τ_χ on price informativeness (Equation 3.24). Rewriting this condition as a lower bound on η :

$$\eta > \eta^* := \frac{\tau_{pub} + 2\tau_\varepsilon + \tau_z}{2\tau_z \Sigma}. \quad (3.33)$$

Taking the same parameters of Figure 3.5 ($k_0 = 0.08$, $\sigma_x = 0.5$, $\rho = 1$, $\tau_\theta = 1$) and moderate values of public precision τ_{pub} , the threshold η^* is on the order of 0.3–0.5. In this calibrated economy, the negative effect of the firm's data availability on FPE requires $N(\tau_{pub})$ to grow by roughly 15–25% for each unit of public information. In other words, each public signal disclosed by a firm should generate between 1.15 and 1.25 units of public information, e.g., news articles. Given the explosion of derivative content around corporate disclosures, this growth rate seems well within plausible magnitudes. In such an economy, the model predicts that more data adoption on the firm side will reduce, not increase, the informativeness of stock prices.

The model also speaks to our understanding of how new technologies affect the interaction between firms and financial markets. The standard view holds that better data at firms produces more informative disclosure, and hence more informative prices and a lower cost of capital. However, an explosion of public information can generate negative effects on investors' ability to generate and search for profitable

signals. Better data at the firm generates more disclosure, more disclosure inflates the volume of data about the firm, and the larger volume raises research costs, which can reduce the informativeness of prices. Whether the net effect on price informativeness is positive or negative depends on how fast public information grows with respect to profitable signals, i.e., signals that investors can use to generate profit.

The results in this paper might have implications for disclosure regulation. For example, mandatory disclosure standards may inflate the pool of public information beyond the level that maximizes the information available to the market. In this case, even mandatory disclosure could reduce the total information available to the market. It is worth noticing that the impact of public disclosures on the market equilibrium hinges also on the amount of additional content that is generated by media, analysts, or other commentary around a firm's disclosure. Moreover, the cross-firm externality in Section 3.3.4 shows that mandating disclosure at one firm redirects research toward others, lowering the non-disclosing firm's cost of capital. These two forces play in opposite directions, and the optimal design might depend on the relative magnitude of η and the cross-firm spillover.

3.5 Conclusion

Firms increasingly adopt technologies that improve their internal information, and what a firm knows affects what it can disclose. This paper studies how this transformation affects the information produced by financial markets. Answering this question requires studying two decisions jointly: the manager's choice of how precisely to disclose, and investors' choice of how much information to produce on their own. The link between the two generates a negative externality of public disclosure on investors' cost of searching for profitable signals. When the firm discloses more, the volume of data available about the firm grows faster than the stock of signals from which investors can profit, and searching for profitable signals becomes more expensive.

The model I study gives two main results. First, technologies that improve the firm's internal information unambiguously increase the manager's optimal disclosure, and even more so when the stock market is noisier. Second, those same technologies can reduce the informativeness of the stock price. When each disclosed signal generates enough derivative content (e.g., media coverage, commentary, social media content), the rising cost of private search shrinks more information than the public disclosure adds. The stock price then reveals less about future fundamentals, when the firm knows more. This reversal cannot arise in standard models with constant

cost of information acquisition, and a calibrated example suggests that it requires a modest amount of derivative content.

The negative externality also extends across firms. When investors allocate limited research capacity across different firms, data adoption at one firm redirects research effort toward its peers, raising their price informativeness and lowering their cost of capital. Part of the market impact of technology adoption therefore appears at firms that did not adopt anything, a benefit that the adopting firm does not capture. The informational consequences of the data economy thus depend not only on what firms learn, but also on how financial markets reorganize their own information production in equilibrium.

The model abstracts from the firm's decision to adopt a data technology and from dynamics. Endogenizing adoption would discipline whether the cross-firm benefit (not internalized by the adopters) leads firms to under-invest in data technologies, and a dynamic version of the model would characterize how the externality evolves as the data economy grows. The model's central prediction is testable. Once equipped with good measures of price informativeness, one could ask whether stock prices have become less informative for firms that gained access to more precise internal data. I plan to pursue these questions in future research.

3.6 Appendix

3.6.1 Model Solution

This appendix reports the step by step solution of the model. As standard in the literature on endogenous information acquisition, I solve the model backward (Veldkamp, 2011).

3.6.1.1 Information Structure and Beliefs

The signal disclosed by the manager is $\omega = s + v = \tilde{\theta} + \chi + v$, with $\chi \sim N(0, \tau_\chi^{-1})$, $v \sim N(0, \tau_v^{-1})$. Conditional on $\tilde{\theta}$, the noise in the public disclosure ω is $u := \chi + v$. Since χ and v are independent, by Bayes' law, the precision of ω about $\tilde{\theta}$ is:

$$\tau_\omega = \left(\frac{1}{\tau_\chi} + \frac{1}{\tau_v} \right)^{-1} = \frac{\tau_\chi \tau_v}{\tau_\chi + \tau_v}, \quad (3.34)$$

which is Equation (3.3) in the text. The precision of the disclosure is the harmonic mean of the two precisions: the public signal can never be more precise than the manager's own information ($\tau_\omega < \tau_\chi$) nor than the disclosure noise ($\tau_\omega < \tau_v$).

Differentiating Equation (3.34):

$$\frac{\partial \tau_\omega}{\partial \tau_\chi} = \frac{\tau_v^2}{(\tau_\chi + \tau_v)^2} > 0, \quad \frac{\partial \tau_\omega}{\partial \tau_v} = \frac{\tau_\chi^2}{(\tau_\chi + \tau_v)^2} > 0. \quad (3.35)$$

Moreover, differentiating the second expression with respect to the firm's internal information (τ_χ), one can show that more precise internal information raises the marginal effect of disclosure precision on the informativeness of the public signal.

$$\frac{\partial^2 \tau_\omega}{\partial \tau_\chi \partial \tau_v} = \frac{2\tau_\chi(\tau_\chi + \tau_v)^2 - 2\tau_\chi^2(\tau_\chi + \tau_v)}{(\tau_\chi + \tau_v)^4} = \frac{2\tau_\chi \tau_v}{(\tau_\chi + \tau_v)^3} > 0. \quad (3.36)$$

In $t = 1$, every investor i observes three signals about $\tilde{\theta}$: (i) the public disclosure ω , (ii) the private signal $y_i = \tilde{\theta} + \varepsilon_i$, and (iii) the price signal $z_P = \tilde{\theta} + (p_x/p_\theta)\tilde{x}$ (derived in Section 3.6.1.3). The noise terms u , ε_i , and $(p_x/p_\theta)\tilde{x}$ are mutually independent, so the three signals are conditionally independent given $\tilde{\theta}$, with precisions τ_ω , τ_ε , and τ_z , respectively. Since all random variables are normally distributed, by Bayes' law, the precisions add and the posterior distribution is given by:

$$\tilde{\theta} \mid \omega, y_i, z_P \sim N\left(\frac{\tau_\omega \omega + \tau_\varepsilon y_i + \tau_z z_P}{\Sigma}, \frac{1}{\Sigma}\right), \quad (3.37)$$

where $\Sigma := \tau_\theta + \tau_\omega + \tau_\varepsilon + \tau_z = \tau_{pub} + \tau_\varepsilon + \tau_z$ is the total market precision defined in the text (recall $\tau_{pub} = \tau_\theta + \tau_\omega$). Without loss of generality, the prior mean is zero so it does not appear in the posterior mean.

3.6.1.2 Portfolio Choice

Each investor i has initial wealth w_0 , buys D_i shares of the risky asset at price $\tilde{P}$, and invests the remainder in the risk-free asset (whose return is normalized to one). Her terminal wealth in $t = 2$ is therefore

$$w_i = w_0 + D_i (\tilde{\theta} - \tilde{P}). \quad (3.38)$$

Denote the posterior moments from Equation (3.37) by $\mu_i := \mathbb{E}[\tilde{\theta} \mid \omega, y_i, z_P]$ and $\sigma_i^2 := \mathbb{V}ar[\tilde{\theta} \mid \omega, y_i, z_P] = 1/\Sigma$. Since terminal wealth is normally distributed (conditional on investor i 's information set), standard derivations of CARA-normal utilities yield:

$$\mathbb{E}[-e^{-\rho w_i} \mid \omega, y_i, z_P] = -\exp\left(-\rho [w_0 + D_i(\mu_i - \tilde{P})] + \frac{\rho^2 D_i^2 \sigma_i^2}{2}\right). \quad (3.39)$$

Maximizing Equation (3.39) over D_i is equivalent to maximizing the mean-variance objective $D_i(\mu_i - \tilde{P}) - (\rho/2)D_i^2\sigma_i^2$. The first-order condition is

$$(\mu_i - \tilde{P}) - \rho D_i \sigma_i^2 = 0 \quad \implies \quad D_i = \frac{\mu_i - \tilde{P}}{\rho \sigma_i^2}, \quad (3.40)$$

and the second-order condition always holds (since $-\rho \sigma_i^2 < 0$). Substituting $\mu_i = (\tau_\omega \omega + \tau_\varepsilon y_i + \tau_z z_P)/\Sigma$ and $\sigma_i^2 = 1/\Sigma$:

$$D_i = \frac{\Sigma}{\rho} (\mu_i - \tilde{P}) = \frac{1}{\rho} \left[\tau_\varepsilon y_i + \tau_\omega \omega + \tau_z z_P - \Sigma \tilde{P} \right], \quad (3.41)$$

which is Equation (3.12) in the main text. As standard in CARA-normal settings, the demand is linear in the signals. Each investor buys when her precision-weighted information exceeds the market price, and her trading aggressiveness is inversely proportional to risk aversion.

3.6.1.3 The Linear Rational-Expectations Equilibrium

Since I am searching for a linear equilibrium, I conjecture and later verify a linear price function:

$$\tilde{P} = p_0 + p_\omega \omega + p_\theta \tilde{\theta} + p_x \tilde{x}. \quad (3.42)$$

Each investor observes $\tilde{P}$ and ω , and knows the equilibrium coefficients. Subtracting the known components and dividing by p_θ isolates the informational content of the price:

$$z_P := \frac{\tilde{P} - p_0 - p_\omega \omega}{p_\theta} = \tilde{\theta} + \frac{p_x}{p_\theta} \tilde{x}. \quad (3.43)$$

The price signal is $\tilde{\theta}$ plus noise $(p_x/p_\theta)\tilde{x} \sim N(0, (p_x/p_\theta)^2 \sigma_x^2)$, so its precision is equal to

$$\tau_z = \frac{p_\theta^2}{p_x^2 \sigma_x^2}. \quad (3.44)$$

In the symmetric equilibrium every investor chooses the same precision, $\tau_{\varepsilon,i} = \tau_\varepsilon$. Integrating the individual demands in Equation (3.41) over the unit mass of investors, the idiosyncratic noise terms average out by the law of large numbers, $\int_0^1 y_i di = \tilde{\theta} + \int_0^1 \varepsilon_i di = \tilde{\theta}$:

$$\int_0^1 D_i di = \frac{1}{\rho} \left[\tau_\varepsilon \tilde{\theta} + \tau_\omega \omega + \tau_z z_P - \Sigma \tilde{P} \right]. \quad (3.45)$$

Imposing $\int_0^1 D_i di + \tilde{x} = Q$ and substituting $z_P = \tilde{\theta} + (p_x/p_\theta)\tilde{x}$ gives:

$$\frac{1}{\rho} \left[(\tau_\varepsilon + \tau_z) \tilde{\theta} + \tau_\omega \omega + \tau_z \frac{p_x}{p_\theta} \tilde{x} - \Sigma \tilde{P} \right] + \tilde{x} = Q. \quad (3.46)$$

Now, solving for the price:

$$\tilde{P} = -\frac{\rho Q}{\Sigma} + \frac{\tau_\omega}{\Sigma} \omega + \frac{\tau_\varepsilon + \tau_z}{\Sigma} \tilde{\theta} + \frac{\rho + \tau_z (p_x/p_\theta)}{\Sigma} \tilde{x}. \quad (3.47)$$

Equation (3.47) must coincide with the conjecture (3.42). Matching term by term we obtain the endogenous coefficients:

$$p_0 = -\frac{\rho Q}{\Sigma}, \quad p_\omega = \frac{\tau_\omega}{\Sigma}, \quad p_\theta = \frac{\tau_\varepsilon + \tau_z}{\Sigma}, \quad p_x = \frac{\rho + \tau_z (p_x/p_\theta)}{\Sigma}. \quad (3.48)$$

However, the last equality (p_x) is a fixed point in the noise-to-fundamental ratio p_x/p_θ . I define for simplicity $\beta := p_x/p_\theta$. The ratio β measures how much noise-trader demand moves the price relative to fundamentals. Dividing p_x by p_θ :

$$\beta = \frac{\rho + \tau_z \beta}{\tau_\varepsilon + \tau_z} \iff \beta (\tau_\varepsilon + \tau_z) - \tau_z \beta = \rho \iff \beta = \frac{\rho}{\tau_\varepsilon}. \quad (3.49)$$

The relative sensitivity of price to noise with respect to fundamentals (β) decreases with τ_ε , because better-informed investors trade more aggressively against noise.

Substituting $\beta = \rho/\tau_\varepsilon$ into Equation (3.44):

$$\tau_z = \frac{1}{\beta^2 \sigma_x^2} = \frac{\tau_\varepsilon^2}{\rho^2 \sigma_x^2}, \quad (3.50)$$

which is Equation (3.14) in the text. Substituting back into p_x in (3.48) gives the closed form reported in Equation (3.13):

$$p_x = \frac{\rho + \tau_z \rho / \tau_\varepsilon}{\Sigma} = \frac{\rho + \tau_\varepsilon / (\rho \sigma_x^2)}{\Sigma}. \quad (3.51)$$

All four coefficients are in terms of $(\tau_{pub}, \tau_\varepsilon, \tau_z)$ and other primitives, with τ_z given by (3.50). The price function is linear as conjectured. Given τ_ε , the linear equilibrium exists and its coefficients are unique. $\square$

3.6.1.4 The Value of Private Information and Optimal Research

I now derive the objective function in Equation (3.15), and then solve for the optimal research precision. These derivation steps are standard in endogenous information acquisition models (Veldkamp, 2011, Chapter 7).

Substituting the optimal demand (3.40) back into Equation (3.39), the exponent becomes $-\rho w_0 - (\mu_i - \tilde{P})^2 / (2\sigma_i^2)$, so the maximized conditional expected utility is

$$\mathbb{E} \left[-e^{-\rho w_i^*} \mid \omega, y_i, z_P \right] = -\exp \left(-\rho w_0 - \frac{(\mu_i - \tilde{P})^2}{2\sigma_i^2} \right). \quad (3.52)$$

Trading gains increase with the squared distance between the investor's expectation and the price, scaled by her posterior variance.

Each investor chooses τ_ε in $t = 1$, before observing y_i , taking the price function and τ_z as given (investors are atomistic). Let $\mathcal{I}_U := \{\omega, z_P\}$ denote the public information set and $\mathcal{I}_i := \{\omega, y_i, z_P\}$ the private one, with conditional variances (from Bayes' law):

$$\sigma_U^2 := \mathbb{V}ar[\tilde{\theta} \mid \mathcal{I}_U] = \frac{1}{\tau_{pub} + \tau_z}, \quad \sigma_i^2 = \frac{1}{\Sigma}. \quad (3.53)$$

I compute the ex-ante expected utility by conditioning on $\mathcal{I}_U$ first. For simplicity, define the trading gain $G := \mu_i - \tilde{P}$ and $G_U := \mathbb{E}[\tilde{\theta} \mid \mathcal{I}_U] - \tilde{P}$ (G_U is the trading gain of an observer with access to public information only). By Bayesian updating, the private (posterior) mean is:

$$\mu_i = \mathbb{E}[\tilde{\theta} \mid \mathcal{I}_U] + \frac{\tau_\varepsilon}{\Sigma} \left(y_i - \mathbb{E}[\tilde{\theta} \mid \mathcal{I}_U] \right). \quad (3.54)$$

Hence, conditional on $\mathcal{I}_U$, G is normal with mean G_U and variance

$$\mathbb{V}ar[G | \mathcal{I}_U] = \mathbb{V}ar[\mu_i | \mathcal{I}_U] = \mathbb{V}ar[\tilde{\theta} | \mathcal{I}_U] - \mathbb{E}[\mathbb{V}ar[\tilde{\theta} | \mathcal{I}_i] | \mathcal{I}_U] = \sigma_U^2 - \sigma_i^2, \quad (3.55)$$

where the second equality is the law of total variance (the conditional variance σ_i^2 is deterministic). From the normality of G , it always holds that (Veldkamp, 2011, see Section 7.1)⁴:

$$\mathbb{E} \left[e^{-G^2/(2\sigma_i^2)} \middle| \mathcal{I}_U \right] = \frac{\sigma_i}{\sigma_U} \exp \left(-\frac{G_U^2}{2\sigma_U^2} \right). \quad (3.56)$$

Taking the unconditional expectation of the expected utility above (Equation (3.52)):

$$\mathbb{E} \left[-e^{-\rho w_i^*} \right] = -e^{-\rho w_0} \frac{\sigma_i}{\sigma_U} \mathbb{E} \left[e^{-G_U^2/(2\sigma_U^2)} \right]. \quad (3.57)$$

A crucial result for solving models with endogenous information acquisition is that the last unconditional expected utility depends only on public information and the price function, not on τ_ε .

Since the CARA-normal certainty equivalent of U is $-\ln(-U)/\rho$, from Equation (3.57) the certainty-equivalent gain from precision τ_ε relative to observing only public information ($\tau_\varepsilon = 0$) is

$$-\frac{1}{\rho} \ln \left(\frac{\sigma_i}{\sigma_U} \right) = \frac{1}{2\rho} \ln \left(\frac{\sigma_U^2}{\sigma_i^2} \right) = \frac{1}{2\rho} \ln \left(\frac{\tau_{pub} + \tau_\varepsilon + \tau_z}{\tau_{pub} + \tau_z} \right), \quad (3.58)$$

which is the benefit term in Equation (3.15). Net of the search cost, the investor now solves

$$\max_{\tau_\varepsilon \geq 0} \frac{1}{2\rho} \ln \left(\frac{\tau_{pub} + \tau_\varepsilon + \tau_z}{\tau_{pub} + \tau_z} \right) - \frac{k(\tau_{pub})}{2} \tau_\varepsilon^2. \quad (3.59)$$

Differentiating with respect to τ_ε , holding τ_z fixed (investors are atomistic and therefore also take price informativeness as given):

$$\frac{1}{2\rho(\tau_{pub} + \tau_\varepsilon + \tau_z)} = k(\tau_{pub}) \tau_\varepsilon \iff \frac{1}{2\rho \Sigma} = k(\tau_{pub}) \tau_\varepsilon, \quad (3.60)$$

which is Equation (3.16) in the main text.

⁴For a normal variable $X \sim N(m, s^2)$ and a constant $a > 0$ it always holds that $\mathbb{E} \left[e^{-G^2/(2\sigma_i^2)} \middle| \mathcal{I}_U \right] = \frac{\sigma_i}{\sigma_U} \exp \left(-\frac{G_U^2}{2\sigma_U^2} \right)$. Substituting $m = G_U$, $s^2 = \sigma_U^2 - \sigma_i^2$, and $a = \sigma_i$ for G , gives (3.56).

3.6.1.5 Comparative Statics

I now derive the response of the equilibrium investors' precisions to a change in public precision τ_{pub} (Equations (3.17), (3.18), and (3.19) in the main text). Recall that τ_ε denotes the equilibrium research precision $\tau_\varepsilon^*(\tau_{pub})$, $\eta = k'(\tau_{pub})/k(\tau_{pub})$ is the cost semi-elasticity from Equation (3.8), and $\Gamma := \tau_{pub} + 2\tau_\varepsilon + 3\tau_z$.

One can rewrite the equilibrium first-order condition (3.60) as

$$2\rho k(\tau_{pub}) \tau_\varepsilon \Sigma = 1, \quad (3.61)$$

with $\Sigma = \tau_{pub} + \tau_\varepsilon + \tau_z$ and $\tau_z = \tau_\varepsilon^2/(\rho^2 \sigma_x^2)$.

Now, differentiating (3.61) with respect to τ_{pub} (notice the right-hand side is constant):

$$2\rho \left[k'(\tau_{pub}) \tau_\varepsilon \Sigma + k(\tau_{pub}) \left(\Sigma \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} + \tau_\varepsilon \frac{\partial \Sigma}{\partial \tau_{pub}} \right) \right] = 0. \quad (3.62)$$

Dividing by $2\rho k(\tau_{pub}) > 0$ and substituting $\eta = k'(\tau_{pub})/k(\tau_{pub})$ gives:

$$\eta \tau_\varepsilon \Sigma + \Sigma \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} + \tau_\varepsilon \frac{\partial \Sigma}{\partial \tau_{pub}} = 0. \quad (3.63)$$

From $\Sigma = \tau_{pub} + \tau_\varepsilon + \tau_z$ and $\tau_z = \tau_\varepsilon^2/(\rho^2 \sigma_x^2)$, it holds that $\partial \tau_z / \partial \tau_\varepsilon = 2\tau_\varepsilon/(\rho^2 \sigma_x^2) = 2\tau_z/\tau_\varepsilon$. Thus:

$$\frac{\partial \Sigma}{\partial \tau_{pub}} = 1 + \left(1 + \frac{2\tau_z}{\tau_\varepsilon} \right) \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}}. \quad (3.64)$$

Substituting (3.64) into (3.63):

$$\eta \tau_\varepsilon \Sigma + \Sigma \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} + \tau_\varepsilon + (\tau_\varepsilon + 2\tau_z) \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} = 0. \quad (3.65)$$

Collecting all the terms in $\partial \tau_\varepsilon / \partial \tau_{pub}$ and since $\Sigma + \tau_\varepsilon + 2\tau_z = \tau_{pub} + 2\tau_\varepsilon + 3\tau_z = \Gamma$:

$$\frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} = -\frac{\tau_\varepsilon (1 + \eta \Sigma)}{\Gamma} < 0, \quad (3.66)$$

which proves Equation (3.17) in the main text.

Similarly, since $\partial \tau_z / \partial \tau_\varepsilon = 2\tau_z/\tau_\varepsilon$ and (3.66):

$$\frac{\partial \tau_z}{\partial \tau_{pub}} = \frac{2\tau_z}{\tau_\varepsilon} \frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} = -\frac{2\tau_z (1 + \eta \Sigma)}{\Gamma} < 0, \quad (3.67)$$

which proves Equation (3.18) in the main text.

Finally, substituting (3.66) into (3.64):

$$\begin{aligned}\frac{\partial \Sigma}{\partial \tau_{pub}} &= 1 - \left(\frac{\tau_\varepsilon + 2\tau_z}{\tau_\varepsilon} \right) \frac{\tau_\varepsilon (1 + \eta \Sigma)}{\Gamma} = \frac{\Gamma - (\tau_\varepsilon + 2\tau_z)(1 + \eta \Sigma)}{\Gamma} \\ &= \frac{\tau_{pub} + \tau_\varepsilon + \tau_z - (\tau_\varepsilon + 2\tau_z) \eta \Sigma}{\Gamma} = \frac{\Sigma [1 - (\tau_\varepsilon + 2\tau_z) \eta]}{\Gamma},\end{aligned}\quad (3.68)$$

which proves Equation (3.19) in the main text. The third equality in the steps above uses $\Gamma - (\tau_\varepsilon + 2\tau_z) = \tau_{pub} + \tau_\varepsilon + \tau_z = \Sigma$.

Therefore, in the symmetric equilibrium:

$$\frac{\partial \tau_\varepsilon}{\partial \tau_{pub}} = -\frac{\tau_\varepsilon (1 + \eta \Sigma)}{\Gamma} < 0, \quad \frac{\partial \tau_z}{\partial \tau_{pub}} = -\frac{2\tau_z (1 + \eta \Sigma)}{\Gamma} < 0, \quad \frac{\partial \Sigma}{\partial \tau_{pub}} = \frac{\Sigma [1 - (\tau_\varepsilon + 2\tau_z) \eta]}{\Gamma}.$$

3.6.1.6 The Manager's Problem

Taking unconditional expectations of the equilibrium price (3.47), and since by construction $\mathbb{E}[\omega] = \mathbb{E}[\tilde{\theta}] = \mathbb{E}[\tilde{x}] = 0$:

$$\mathbb{E}[\tilde{P}] = p_0 = -\frac{\rho Q}{\Sigma}. \quad (3.69)$$

The expected price equals the (negative of the) risk premium. The stock price, in expectation, rises with total market precision Σ because more informed investors require lower compensation for risk.

The manager chooses the precision of her public disclosure τ_v in $t = 0$ to maximize $U_M = p_0 - C(\tau_v)$. The manager knows that Σ responds to τ_{pub} through the market equilibrium of Section 3.6.1.5. The marginal benefit of the manager with respect to changes in her public disclosure precision is

$$\text{MB}(\tau_v) := \frac{\partial p_0}{\partial \tau_v} = \frac{\rho Q}{\Sigma^2} \frac{\partial \Sigma}{\partial \tau_{pub}} \frac{\partial \tau_{pub}}{\partial \tau_v}, \quad (3.70)$$

where, from $\tau_{pub} = \tau_\theta + \tau_\chi \tau_v / (\tau_\chi + \tau_v)$ and Equation (3.35),

$$\frac{\partial \tau_{pub}}{\partial \tau_v} = \frac{\tau_\chi^2}{(\tau_\chi + \tau_v)^2} > 0. \quad (3.71)$$

The sign of the marginal benefit is therefore the sign of $\partial \Sigma / \partial \tau_{pub}$ (see below).

Proof of Proposition 3.1 (Optimal Disclosure). The manager's first-order condition is $\text{MB}(\tau_v^*) = C'(\tau_v^*)$. I verify that the net marginal benefit $\text{MB}(\tau_v) - C'(\tau_v)$ is strictly positive at $\tau_v = 0$ and strictly negative for τ_v large, so an interior solution exists.

At $\tau_v = 0$: (i) $\partial\tau_{pub}/\partial\tau_v = \tau_\chi^2/(\tau_\chi + 0)^2 = 1$; (ii) $C'(0) = 0$ by convexity of C with $C'(0) = 0$; and (iii) $\tau_{pub} = \tau_\theta$, so by the assumption of the proposition, $\partial\Sigma/\partial\tau_{pub} > 0$ at $\tau_{pub} = \tau_\theta$, i.e., $(\tau_\varepsilon + 2\tau_z)\eta < 1$ at that point. Therefore $MB(0) = (\rho Q/\Sigma^2)(\partial\Sigma/\partial\tau_{pub}) > 0 = C'(0)$. Under the proposition's assumption ($\partial\Sigma/\partial\tau_{pub} > 0$ at the starting point $\tau_{pub} = \tau_\theta$, i.e., $\tau_v = 0$), silence is never optimal.

As $\tau_v \rightarrow \infty$, $\partial\tau_{pub}/\partial\tau_v = \tau_\chi^2/(\tau_\chi + \tau_v)^2 \rightarrow 0$, while $\partial\Sigma/\partial\tau_{pub}$ and $\rho Q/\Sigma^2$ are bounded (as $\tau_{pub} \rightarrow \tau_\theta + \tau_\chi < \infty$). Thus: $MB(\tau_v) \rightarrow 0$. At the same time, $C'(\tau_v)$ is strictly increasing, so $C'(\tau_v) > 0$ bounded away from zero for large τ_v . The net marginal benefit is eventually strictly negative.

Using the steps above, $MB - C'$ crosses zero at some $\tau_v^* \in (0, \infty)$. Strict quasi-concavity of U_M in τ_v —which holds whenever C is sufficiently convex—implies the crossing is unique, and τ_v^* is the global maximum. $\square$

For studying how optimal disclosure interacts with noise in the stock price, it is useful to define the net marginal benefit as

$$H(\tau_v; \tau_\chi, \sigma_x) := MB(\tau_v; \tau_\chi, \sigma_x) - C'(\tau_v), \quad (3.72)$$

so the interior optimum satisfies $H(\tau_v^*; \cdot) = 0$.

By the implicit function theorem at the interior optimum ($H(\tau_v^*; \sigma_x) = 0$), one can study how optimal disclosure changes with stock market noise:

$$\frac{d\tau_v^*}{d\sigma_x} = -\frac{\partial H/\partial\sigma_x}{\partial H/\partial\tau_v}, \quad (3.73)$$

with $\partial H/\partial\tau_v < 0$, so the sign of $(d\tau_v^*/d\sigma_x)$ is determined by the sign of $(\partial MB/\partial\sigma_x)$. The stock market noise σ_x enters the marginal benefit through the market equilibrium. Higher σ_x mechanically reduces the precision of the price signal and this direct effect dominates the equilibrium response of τ_ε , so total market precision $\Sigma = \tau_{pub} + \tau_\varepsilon + \tau_z$ falls. A lower Σ raises the risk-premium required by investors ($\rho Q/\Sigma^2$). When the market learns little from the price, the cost of capital is high and each unit of public precision buys a larger reduction in the risk premium. This mechanism raises the manager's marginal benefit in disclosing public information. Therefore, optimal disclosure increases with stock-market noise, $d\tau_v^*/d\sigma_x > 0$, whenever $\partial MB/\partial\sigma_x > 0$.

Moreover, the sensitivity of optimal disclosure to the firm's internal information is stronger when there is more noise in the stock price, i.e., $d\tau_v^*/d\tau_\chi$ increases in σ_x . By the implicit function theorem, $d\tau_v^*/d\tau_\chi = -H_{\tau_\chi}/H_{\tau_v}$, with $H_{\tau_\chi} = \partial MB/\partial\tau_\chi > 0$ and $H_{\tau_v} < 0$. The numerator H_{τ_χ} is proportional to risk-premium required by investors $\rho Q/\Sigma^2$. When σ_x is high, $\tau_z = \tau_\varepsilon^2/(\rho^2\sigma_x^2)$ is small, Σ is low, and $\rho Q/\Sigma^2$ is

large. Therefore, the same increase in the disclosure precision translates into a larger increase in the marginal benefit. When σ_x is low, the opposite holds and H_{τ_x} is small.

3.6.1.7 Price Informativeness

Proof of Proposition 3.2 (Price Informativeness). Recall that price informativeness is defined as $\Phi = \tau_{pub} + \tau_z$. Using Equation (3.18)

$$\frac{\partial \Phi}{\partial \tau_{pub}} = 1 + \frac{\partial \tau_z}{\partial \tau_{pub}} = 1 - \frac{2\tau_z(1 + \eta \Sigma)}{\Gamma} = \frac{\Gamma - 2\tau_z(1 + \eta \Sigma)}{\Gamma}. \quad (3.74)$$

Now, since $\Gamma = \tau_{pub} + 2\tau_\varepsilon + 3\tau_z$:

$$\Gamma - 2\tau_z(1 + \eta \Sigma) = \tau_{pub} + 2\tau_\varepsilon + \tau_z - 2\tau_z \eta \Sigma, \quad (3.75)$$

which delivers Equation (3.23) in the text.

From (3.75), $\partial \Phi / \partial \tau_{pub} < 0$ if and only if the condition $2\tau_z \eta \Sigma > \tau_{pub} + 2\tau_\varepsilon + \tau_z$ holds. When $2\eta \Sigma > 1$, rearranging gives the equivalent threshold in the main text, $\tau_z > (\tau_{pub} + 2\tau_\varepsilon) / (2\eta \Sigma - 1)$. When $2\eta \Sigma \leq 1$, the condition above does not hold and informativeness always increases. The firm's internal information affects Φ only through public disclosure τ_{pub} :

$$\frac{\partial \Phi}{\partial \tau_x} = \frac{\partial \Phi}{\partial \tau_{pub}} \cdot \frac{\partial \tau_{pub}}{\partial \tau_x}, \quad \frac{\partial \tau_{pub}}{\partial \tau_x} = \frac{\tau_v^2}{(\tau_x + \tau_v)^2} > 0 \text{ for } \tau_v > 0, \quad (3.76)$$

using Equation (3.35). Hence price informativeness decreases with the firm's internal data when the condition above holds. $\square$

3.6.1.8 The Two-Firm Economy

With independent fundamentals, the certainty-equivalent gains from the two research precisions are additive (the derivation of Section 3.6.1.4 applies firm by firm), while the search cost is joint:

$$\max_{\tau_{\varepsilon,A}, \tau_{\varepsilon,B} \geq 0} \sum_{i \in \{A,B\}} \frac{1}{2\rho} \ln \left(\frac{\Sigma_i}{\tau_{pub,i} + \tau_{z,i}} \right) - \frac{k_0}{2} (r_A \tau_{\varepsilon,A} + r_B \tau_{\varepsilon,B})^2. \quad (3.77)$$

Define the total search effort $E := r_A \tau_{\varepsilon,A} + r_B \tau_{\varepsilon,B}$. The two first-order conditions (holding $\tau_{z,i}$ fixed) are

$$\frac{1}{2\rho \Sigma_A} = k_0 E r_A, \quad \frac{1}{2\rho \Sigma_B} = k_0 E r_B. \quad (3.78)$$

Each precision is produced until its marginal value equals its marginal search cost. The joint cost implies that the marginal cost of researching firm i is proportional to the *total* effort E , scaled by firm i 's r_i . Dividing the first condition by the second:

$$\frac{\Sigma_B}{\Sigma_A} = \frac{r_A}{r_B} \iff r_A \Sigma_A = r_B \Sigma_B, \quad (3.79)$$

which is the condition in Equation (3.26). In equilibrium, the cost-adjusted total precision must be the same across firms.

Proof of Proposition 3.3 (Cross-Firm Spillover). From the symmetric equilibrium: $\tau_{pub,A} = \tau_{pub,B} = \tau_{pub}$, $r_A = r_B = r$, $\tau_{\varepsilon,A} = \tau_{\varepsilon,B} = \tau_{\varepsilon}$, $\Sigma_A = \Sigma_B = \Sigma$, so $E = 2r\tau_{\varepsilon}$ and $M = r\Sigma$. An increase in $\tau_{\chi,A}$ raises $\tau_{pub,A}$ through $\partial\tau_{pub,A}/\partial\tau_{\chi,A} = \tau_v^2/(\tau_{\chi} + \tau_v)^2 > 0$. This holds with disclosure fixed, the endogenous disclosure response would only reinforce the increase in $\tau_{pub,A}$. It is therefore sufficient to study the effects of $d\tau_{pub,A} > 0$. Denoting for simplicity $a := 1 + 2\tau_z/\tau_{\varepsilon} > 0$, I can rewrite $d\Sigma_i = d\tau_{pub,i} + a d\tau_{\varepsilon,i}$.

The equilibrium is characterized by conditions (3.79) and (3.85) (see below). Taking the total derivative of (3.79), with $dr_A = r' d\tau_{pub,A}$ and $dr_B = d\tau_{pub,B} = 0$:

$$r' \Sigma d\tau_{pub,A} + r (d\tau_{pub,A} + a d\tau_{\varepsilon,A}) = r a d\tau_{\varepsilon,B} \iff r a (d\tau_{\varepsilon,B} - d\tau_{\varepsilon,A}) = (r' \Sigma + r) d\tau_{pub,A}. \quad (3.80)$$

Now, differentiating (3.85) (the right-hand side is constant), $M dE + E dM = 0$, with $dE = r' \tau_{\varepsilon} d\tau_{pub,A} + r d\tau_{\varepsilon,A} + r d\tau_{\varepsilon,B}$ and $dM = d(r_B \Sigma_B) = r a d\tau_{\varepsilon,B}$:

$$M [r' \tau_{\varepsilon} d\tau_{pub,A} + r d\tau_{\varepsilon,A} + r d\tau_{\varepsilon,B}] + E r a d\tau_{\varepsilon,B} = 0. \quad (3.81)$$

Substituting $d\tau_{\varepsilon,A} = d\tau_{\varepsilon,B} - \frac{r' \Sigma + r}{r a} d\tau_{pub,A}$ from (3.80) into (3.81) and collecting terms:

$$\frac{d\tau_{\varepsilon,B}}{d\tau_{pub,A}} = \frac{M [r' \Sigma + r - a r' \tau_{\varepsilon}]}{a (2Mr + Era)} = \frac{Mr \left[1 + \frac{\eta}{2} (\tau_{pub} - \tau_z) \right]}{a (2Mr + Era)}, \quad (3.82)$$

where the second equality uses $\Sigma - a \tau_{\varepsilon} = \Sigma - \tau_{\varepsilon} - 2\tau_z = \tau_{pub} - \tau_z$ and $r' = (\eta/2) r$ (from $\eta = 2r'/r$). The denominator is always strictly positive. The numerator is strictly positive when $\eta (\tau_z - \tau_{pub}) < 2$, and this condition always holds when public information raises total market precision. $(\tau_{\varepsilon} + 2\tau_z) \eta < 1$ implies $\eta \tau_z < 1/2$, so $\eta (\tau_z - \tau_{pub}) < 1/2 < 2$. Therefore $d\tau_{\varepsilon,B}/d\tau_{pub,A} > 0$: investors reallocate research effort toward firm B .

Substituting (3.82) back into (3.80):

$$\frac{d\tau_{\varepsilon,A}}{d\tau_{pub,A}} = - \frac{r M r' \Gamma + Mr^2 + E a (r' \Sigma + r)}{a r (2Mr + Era)} < 0, \quad (3.83)$$

where the simplification of the numerator uses $\tau_{pub} - \tau_z - 2\Sigma = -(\tau_{pub} + 2\tau_\varepsilon + 3\tau_z) = -\Gamma$. Every term in the numerator is positive, so research effort on firm A decreases.

Finally, since $\tau_{pub,B}$ is fixed, $d\Sigma_B = a d\tau_{\varepsilon,B} > 0$ and, from $\tau_{z,B} = \tau_{\varepsilon,B}^2 / (\rho^2 \sigma_x^2)$, $d\tau_{z,B} = (2\tau_z / \tau_\varepsilon) d\tau_{\varepsilon,B} > 0$. Hence firm B 's price informativeness increases, $d\Phi_B = d\tau_{z,B} > 0$, and its price rises (in expectation):

$$d\mathbb{E}[\tilde{P}_B] = \frac{\rho Q}{\Sigma_B^2} d\Sigma_B > 0, \quad (3.84)$$

i.e., firm B 's cost of capital falls. □

Notice also that, multiplying the condition in (3.78) by the common value $M := r_A \Sigma_A = r_B \Sigma_B$ pins down total effort:

$$E M = \frac{1}{2\rho k_0}. \quad (3.85)$$

3.6.2 Extension with Sequential Search: Derivations

This appendix reports the derivations for the extension in Section 3.4 with sequential search. This extension builds on [Menzio \(2007\)](#) directed search. I first derive the search cost in Equation (3.29), then the equilibrium ratio between profitable signals and total number of signals under partially directed search. Everything else in the model remains unchanged. Finally, I show that the externality of public information and the paper's main results hold whenever the screening technology is imperfect ($\delta < 1$).

3.6.2.1 The Sequential Search Cost

The data environment consists of $N(\tau_{pub})$ total data points, n_0 are informative (each with precision τ_d about $\tilde{\theta}$ beyond ω) and $N(\tau_{pub}) - n_0$ do not have any additional precision with respect to the information known by the public (e.g., they are redundant). The density of profitable signals is $q(\tau_{pub}) = n_0/N(\tau_{pub})$ and the signal-to-noise ratio is $r(\tau_{pub}) = 1/q(\tau_{pub}) = N(\tau_{pub})/n_0$. There are several content creators who enter the sector freely and compete. The number of redundant data points is pinned down by the free-entry condition (3.27) in this sector. Since $\partial\pi/\partial N < 0$ and $\partial\pi/\partial\tau_{pub} > 0$, the total amount of data points changes with the precision of the public signal:

$$N'(\tau_{pub}) = -\frac{\partial\pi/\partial\tau_{pub}}{\partial\pi/\partial N} > 0, \quad (3.86)$$

so the pool of raw data increases with public precision. The amount of informative data points n_0 is fixed, thus $r'(\tau_{pub}) = N'(\tau_{pub})/n_0 > 0$.

Investors are Bayesian and all signals are independent, so the total precision of m informative data points is $m \tau_d$. To reach precision τ_ε , the investor must therefore find $m = \tau_\varepsilon/\tau_d$ profitable data points. Under random sequential search (as in the baseline model), each draw is informative with probability $q(\tau_{pub})$, so the number of draws needed to find one informative data point is $1/q(\tau_{pub})$. The expected number of draws needed to find m informative data points is

$$S = \frac{m}{q(\tau_{pub})} = \frac{\tau_\varepsilon}{\tau_d q(\tau_{pub})} = \frac{\tau_\varepsilon r(\tau_{pub})}{\tau_d}, \quad (3.87)$$

which is Equation (3.28) in the text. Processing S data points costs $(c/2) S^2$, so the

total cost of achieving precision τ_ε is

$$\frac{c}{2} S^2 = \frac{c}{2} \left(\frac{\tau_\varepsilon r(\tau_{pub})}{\tau_d} \right)^2 = \frac{c r(\tau_{pub})^2}{2 \tau_d^2} \tau_\varepsilon^2 = \frac{k(\tau_{pub})}{2} \tau_\varepsilon^2, \quad (3.88)$$

where, for simplicity, I denoted $k(\tau_{pub}) = k_0 r(\tau_{pub})^2$ and $k_0 = c/\tau_d^2$. This is the cost function in Equation (3.7), and it delivers Equation (3.29) in the main text.

Notice that the “base” cost $k_0 = c/\tau_d^2$ includes the processing cost per draw, and the precision of each informative data point. When each individual (profitable) dataset is noisier (lower τ_d), investors need more draws to reach the same precision, which raises the base cost. Finally, since $r'(\tau_{pub}) > 0$ from Equation (3.86), the cost semi-elasticity η under random search is always positive:

$$\eta = \frac{k'(\tau_{pub})}{k(\tau_{pub})} = \frac{2 r'(\tau_{pub})}{r(\tau_{pub})} = \frac{2 N'(\tau_{pub})}{N(\tau_{pub})} > 0. \quad (3.89)$$

In other words, public precision attracts the entry of content creators (e.g., media, analysts), reduces the density of profitable signals, and forces investors to search longer at an increasing marginal cost.

3.6.2.2 The Effective Signal-to-Noise Ratio

When investors have access to a screening technology with accuracy $\delta \in [1/2, 1]$, each data point is independently flagged as profitable or not before acquiring the signal. The screening technology allows to correctly flag an informative as “likely useful” with probability δ , while a redundant one is incorrectly flagged with probability $1 - \delta$. The investor restricts her sequential search to the pool of data points that are flagged as likely useful. Thus, the screened pool contains $n_0 \delta$ true positives and only $(N(\tau_{pub}) - n_0)(1 - \delta)$ false positives, for a total of

$$P(\tau_{pub}, \delta) = n_0 \delta + (N(\tau_{pub}) - n_0)(1 - \delta). \quad (3.90)$$

With the screening technology, the effective signal density is:

$$q_{\text{eff}}(\tau_{pub}, \delta) = \frac{n_0 \delta}{P(\tau_{pub}, \delta)} = \frac{n_0 \delta}{n_0 \delta + (N(\tau_{pub}) - n_0)(1 - \delta)}, \quad (3.91)$$

which is Equation (3.30) in the text. Taking the inverse and using $N(\tau_{pub}) - n_0 = n_0(r(\tau_{pub}) - 1)$:

$$r_{\text{eff}}(\tau_{pub}, \delta) = \frac{1}{q_{\text{eff}}} = 1 + \frac{(N(\tau_{pub}) - n_0)(1 - \delta)}{n_0 \delta} = 1 + (r(\tau_{pub}) - 1) \frac{1 - \delta}{\delta}, \quad (3.92)$$

which is Equation (3.31) in the text. When $\delta = 1/2$, $(1 - \delta)/\delta = 1$ and $r_{\text{eff}} = r(\tau_{pub})$, the screening technology is uninformative and the model boils down to the baseline case (random search). When $\delta = 1$, $(1 - \delta)/\delta = 0$ and $r_{\text{eff}} = 1$, the screen is perfect and the investor inspects only informative data points. The derivation in Section 3.6.2.1 is unchanged with r_{eff} in place of r . The cost semi-elasticity under partially directed search follows from the derivative of the cost function with respect to τ_{pub} . From Equation (3.92), $\partial r_{\text{eff}} / \partial \tau_{pub} = r'(\tau_{pub}) (1 - \delta) / \delta$, so

$$\eta(\delta) = \frac{\partial k / \partial \tau_{pub}}{k} = \frac{2 \partial r_{\text{eff}} / \partial \tau_{pub}}{r_{\text{eff}}} = \frac{2 r'(\tau_{pub}) (1 - \delta) / \delta}{r_{\text{eff}}(\tau_{pub}, \delta)}. \quad (3.93)$$

3.6.2.3 Discussion of The Externality of Public Information

In this Section, I discuss the three properties of $\eta(\delta)$ discussed in Section 3.4: (i) the externality of public information holds for any imperfect screen, $\eta(\delta) > 0$ for $\delta < 1$; (ii) better screening reduces the effect of the externality; and (iii) in environments with large volumes of data $\eta(\delta)$ converges to the baseline model (random search), so screening raises the threshold for the decline in price informativeness but does not avoid it.

The first property is trivial. Recall from Equation (3.86) that free entry implies $r'(\tau_{pub}) > 0$. For any $\delta < 1$, $(1 - \delta)/\delta > 0$ and $r_{\text{eff}}(\tau_{pub}, \delta) \geq 1 > 0$, so every term in Equation (3.93) is strictly positive and $\eta(\delta) > 0$. As long as investors must inspect a data point to learn its quality, public information raises the search cost.

For the second property, write $\eta(\delta) = 2 r'(\tau_{pub}) g(\delta)$, where

$$g(\delta) := \frac{(1 - \delta)/\delta}{r_{\text{eff}}(\tau_{pub}, \delta)} = \frac{1 - \delta}{\delta + A(1 - \delta)}, \quad A := r(\tau_{pub}) - 1 \geq 0. \quad (3.94)$$

Differentiating with respect to δ :

$$\begin{aligned} g'(\delta) &= \frac{-[\delta + A(1 - \delta)] - (1 - \delta)(1 - A)}{[\delta + A(1 - \delta)]^2} \\ &= \frac{-\delta - A(1 - \delta) - (1 - \delta) + A(1 - \delta)}{[\delta + A(1 - \delta)]^2} = \frac{-1}{[\delta + A(1 - \delta)]^2} < 0. \end{aligned} \quad (3.95)$$

Since $r'(\tau_{pub}) > 0$, $\partial\eta/\partial\delta = 2r'(\tau_{pub})g'(\delta) < 0$ for all $\delta \in [1/2, 1)$. A better screening technology reduces the externality of public information. At the boundary $\delta = 1$, $g(1) = 0$ and $\eta(1) = 0$, with perfect screening the search cost does not depend on τ_{pub} and the model converges to traditional models with fixed cost.

For the third property, recall from Equation (3.33) that price informativeness declines in the firm's internal data when $\eta(\delta) > \eta^* = (\tau_{pub} + 2\tau_\varepsilon + \tau_z)/(2\tau_z\Sigma)$. For firms with large volumes of information available ($r(\tau_{pub}) \gg 1$), the term $(r(\tau_{pub}) - 1)(1 - \delta)/\delta$ dominates the constant in Equation (3.92), so $r_{\text{eff}} \approx r(\tau_{pub})(1 - \delta)/\delta$. Substituting into Equation (3.93), the factor $(1 - \delta)/\delta$ cancels out

$$\eta(\delta) \approx \frac{2r'(\tau_{pub})(1 - \delta)/\delta}{r(\tau_{pub})(1 - \delta)/\delta} = \frac{2r'(\tau_{pub})}{r(\tau_{pub})} = \frac{2N'(\tau_{pub})}{N(\tau_{pub})}, \quad (3.96)$$

which gives the baseline random search η in Equation (3.89). Even when the false positive rate $1 - \delta$ is small, the absolute number of false positives $(N(\tau_{pub}) - n_0)(1 - \delta)$ grows with $N(\tau_{pub})$ and eventually overwhelms the fixed stock of true positives $n_0\delta$ in the screened pool.

References

- Abis, Simona (2022) "Man vs. Machine: Quantitative and Discretionary Equity Management," *Working Paper*.
- Abis, Simona and Laura Veldkamp (2024) "The Changing Economics of Knowledge Production," *Review of Financial Studies*, Vol. 37, No. 5, pp. 1556–1592.
- Admati, Anat R. and Paul Pfleiderer (2000) "Forcing Firms to Talk: Financial Disclosure Regulation and Externalities," *Review of Financial Studies*, Vol. 13, No. 3, pp. 479–519.
- Blankespoor, Elizabeth, Ed de Haan, and Edward Xuejun Li (2026) "AI-Generated Disclosure," *Working Paper*.
- Bond, Philip, Alex Edmans, and Itay Goldstein (2012) "The Real Effects of Financial Markets," *Annual Review of Financial Economics*, Vol. 4, pp. 339–360.
- Bonelli, Maxime (2025) "Data-Driven Venture Capital," *Working Paper*.
- Bonelli, Maxime and Thierry Foucault (2023) "The Rise of Data Analytics and Stock Picking," *Working Paper*.
- Cao, Sean, Kai Du, Baozhong Yang, and Alan L. Zhang (2021) "Copycat Skills and Disclosure Costs: Evidence from Peer Firms' Digital Footprints," *Journal of Accounting Research*, Vol. 59, No. 4, pp. 1261–1302.
- Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L. Zhang (2023) "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI," *Review of Financial Studies*, Vol. 36, No. 9, pp. 3603–3642.
- (2024) "From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses," *Journal of Financial Economics*, Vol. 160, p. 103916.
- Cao, Sean, Grace Ma, Jennifer Wu Tucker, and Chi Wan (2018) "Technological Peer Pressure and Product Disclosure," *The Accounting Review*, Vol. 93, No. 6, pp. 95–126.
- Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng (2025) "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts," *Management Science*, Vol. 71, No. 6, pp. 4599–4621.
- Cong, Lin William, Danxia Xie, and Longtian Zhang (2021) "Knowledge Accumulation, Privacy, and Growth in a Data Economy," *Management Science*, Vol. 67, No. 10, pp. 6480–6492.
- de Haan, Ed, Charles M. C. Lee, Tianyu Liu, and Suzie Noh (2025) "The Shadow Price of Public Information," *Working Paper*.
- Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2024) "Does Alternative Data Improve Financial Forecasting? The Horizon Effect," *Journal of Finance*, Vol. 79, No. 3, pp. 2117–2154.
- Diamond, Douglas W. (1985) "Optimal Release of Information by Firms," *Journal of Finance*, Vol. 40, No. 4, pp. 1071–1094.

- Diamond, Douglas W. and Robert E. Verrecchia (1991) "Disclosure, Liquidity, and the Cost of Capital," *Journal of Finance*, Vol. 46, No. 4, pp. 1325–1359.
- Dugast, Jérôme and Thierry Foucault (2018) "Data Abundance and Asset Price Informativeness," *Journal of Financial Economics*, Vol. 130, No. 2, pp. 367–391.
- (2025) "Equilibrium Data Mining and Data Abundance," *Journal of Finance*, Vol. 80, No. 1, pp. 211–258.
- Dye, Ronald A. (1985) "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, Vol. 23, No. 1, pp. 123–145.
- Farboodi, Maryam and Laura Veldkamp (2020) "Long-Run Growth of Financial Data Technology," *American Economic Review*, Vol. 110, No. 8, pp. 2485–2523.
- (2023) "Data and Markets," *Annual Review of Economics*, Vol. 15, pp. 23–40.
- (2025) "A Model of the Data Economy," *Review of Economic Studies*, forthcoming.
- Frésard, Laurent and Marco Zanolini (2026) "Are New Technologies Replacing the Information Produced by Financial Markets?" *Working Paper*.
- Gao, Pingyang and Pierre Jinghong Liang (2013) "Informational Feedback, Adverse Selection, and Optimal Disclosure Policy," *Journal of Accounting Research*, Vol. 51, No. 5, pp. 1133–1158.
- Goldstein, Itay and Liyan Yang (2017a) "Information Disclosure in Financial Markets," *Annual Review of Financial Economics*, Vol. 9, pp. 101–125.
- (2017b) "Information Disclosure, Real Investment, and Shareholder Welfare," *Journal of Finance*, Vol. 72, No. 1, pp. 197–236.
- Grossman, Sanford J. and Joseph E. Stiglitz (1980) "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, Vol. 70, No. 3, pp. 393–408.
- Hellwig, Martin F. (1980) "On the Aggregation of Information in Competitive Markets," *Journal of Economic Theory*, Vol. 22, No. 3, pp. 477–498.
- Hirshleifer, David and Siew Hong Teoh (2003) "Limited Attention, Information Disclosure, and Financial Reporting," *Journal of Accounting and Economics*, Vol. 36, No. 1–3, pp. 337–386.
- Huang, Shiyang, Yan Xiong, and Liyan Yang (2022) "Skill Acquisition and Data Sales," *Management Science*, Vol. 68, No. 8, pp. 6116–6144.
- 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.
- Menzio, Guido (2007) "A Theory of Partially Directed Search," *Journal of Political Economy*, Vol. 115, No. 5, pp. 748–769.
- Menzio, Guido and Nicholas Trachter (2015) "Equilibrium Price Dispersion with Sequential Search," *Journal of Economic Theory*, Vol. 160, pp. 188–215.
- Peng, Lin and Wei Xiong (2006) "Investor Attention, Overconfidence and Category Learning," *Journal of Financial Economics*, Vol. 80, No. 3, pp. 563–602.

- Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan L. Zhang (2025) "Generative AI and Asset Management," *Working Paper*.
- Sims, Christopher A. (2003) "Implications of Rational Inattention," *Journal of Monetary Economics*, Vol. 50, No. 3, pp. 665–690.
- Tookes, Heather E. (2008) "Information, Trading, and Product Market Interactions: Cross-Sectional Implications of Informed Trading," *Journal of Finance*, Vol. 63, No. 1, pp. 379–413.
- Veldkamp, Laura (2011) *Information Choice in Macroeconomics and Finance*, Princeton, NJ: Princeton University Press.
- Verrecchia, Robert E. (1982) "Information Acquisition in a Noisy Rational Expectations Economy," *Econometrica*, Vol. 50, No. 6, pp. 1415–1430.
- (1983) "Discretionary Disclosure," *Journal of Accounting and Economics*, Vol. 5, pp. 179–194.
- (2001) "Essays on Disclosure," *Journal of Accounting and Economics*, Vol. 32, No. 1–3, pp. 97–180.
- Zanotti, Marco (2026) "Asset (and Data) Managers," *Working Paper*.