The Macroeconomics of Data: Scale, Product Choice, and Pricing in the Information Age*

Vladimir Asriyan Alexandre Kohlhas

March 2026

Abstract

We document a substantial rise in the accuracy of U.S. firms' expectations since the early 2000s, closely linked to firm-size dynamics and consistent with major advances in data-processing technologies. To study the macroeconomic implications, we develop a quantitative framework where firms produce information to optimize their *scale*, *product choice*, and *pricing strategies*. While information enhances firms' productive efficiency, it also facilitates price discrimination, driving a wedge between private and social returns. Calibrating our model to U.S. firm-level data, we find that data-processing advances have significantly increased TFP over the past two decades (3.8–4.4%) by improving firms' scale and product choices. Yet, the associated welfare benefits have been modest (0.3–1.4%), reflecting in part excessive information production. Our findings underscore a central role for data regulation in the modern information economy.

JEL codes: E10, E60, C53, D83, D84

Keywords: data economy, expectations, information frictions, product choice, price discrimination, rent extraction, misallocation, data regulation

*First draft: June 2024. Asriyan: CREI, ICREA, BSE, UPF (email: vasriyan@crei.cat). Kohlhas: University of Oxford (email: alexandre.kohlhas@economics.ox.ac.uk). We are thankful to Isaac Baley, Timo Boppart, Joel David, Jan Eeckhout, William Fuchs, Joachim Hubmer, Guido Lorenzoni, Alberto Martin, Guillermo Ordonez, Guangyu Pei, Giacomo Ponzetto, Edouard Schaal, Laura Veldkamp, Venky Venkateswaran, Jaume Ventura, Ansgar Walther, Joshua Weiss, and seminar and conference participants at CREI, UCSD, USC, University of Chicago, HKU, CUHK, HKUST, ESSEC, Oxford University, Goethe University, NHH, University of Luxembourg, St. Louis Fed, NBER SI Micro Data and Macro Models, UT Austin Macro/Finance Conference, BSE Summer Forum 2024, SED Buenos Aires, the Annual Meetings of the Armenian Economic Association 2023 and of the AEA 2024 Meetings for their feedback and comments. Financial support from the European Research Council (MACFRIC, 101229916) and Handelsbank Stiftelsen is gratefully acknowledged. Erfan Ghofrani provided excellent research assistance.

1 Introduction

Advances in data-processing technologies are widely believed to have the potential to transform economic interactions. Consistent with this view, the past two decades have seen a substantial rise in the share of firms that systematically use data to inform their economic decision-making. A simple estimate based on survey data from a sample of medium-to-large firms indicates that this share has more than doubled in the past ten years alone (Mckinsey and Company, 2023). Recent estimates suggest that approximately 40-75 percent of manufacturing firms employ some form of *data-driven decision-making* (Brynjolfsson and McElheran, 2024).[1]

Despite the substantial rise in data use by firms, the macroeconomic consequences of these developments are, nevertheless, not fully understood. Has the apparent rise in data use led to an improved allocation of resources across and within firms, substantial enough to affect economy-wide dynamics? Or, have these developments primarily enabled firms to extract larger rents from consumers at the expense of social efficiency? Answers to these questions are important not only to understand the developments of the recent past but also to design effective regulatory frameworks that will govern the data economy of the future.

In this paper, we explore answers to these questions. Throughout, we view *data as information* that helps firms predict economic fundamentals (Baley and Veldkamp, 2025). The rationale for this view is both *practical*—changes in information can be proxied by changes in the accuracy of firms' expectations—and *conceptual*—since the seminal work of Lucas (1972), the role of information has been shown central for our understanding of macroeconomic dynamics.[2] Using micro-level data on firms' expectations, we document a substantial rise in the accuracy of U.S. firms' expectations since the early 2000s. This improvement is closely linked to firm-size dynamics and aligns with a significant increase in firms' data use.

Our main contribution is to develop the first *unifying* quantitative framework to study the macroeconomic consequences of these developments. In our model, information production allows firms to optimally choose their *scale*, *product choice*, and *pricing* strategies—three margins widely viewed as central to how data and information technologies have transformed firm behavior over the past decades.[3] In our framework, information has a *dual role*. On the one hand, it enables firms to align their production decisions with underlying fundamentals, improving the efficiency of resource allocation within and across firms. On the other hand, by allowing firms to better segment consumers and tailor prices, it also facilitates price discrim-

[1] See Brynjolfsson and McElheran (2016) and Goldfarb *et al.* (2015) for overviews on adoption rates.

[2] See, e.g., Woodford (2002), Sims (2003), Blanchard *et al.* (2013), Chahrour and Jurado (2018), Angeletos and Lian (2016), Angeletos *et al.* (2021), among others.

[3] Feng *et al.* (2020) and Babina *et al.* (2024) discuss changes to product design, while Adams *et al.* (2025) documents changes in firms' pricing; Zolas *et al.* (2021) and Jin and McElheran (2024) discuss changes in inputs associated with cloud computing and AI; lastly, O'Neill (2023) presents several business cases.

ination, which can distort consumption allocations away from efficiency and alter the social value of data. The aggregate effects of advances in data-processing technologies—and, in turn, the design and desirability of corrective policies—depend on the interaction of these forces.

To quantify the strength of these forces, we calibrate the model using U.S. firm-level data to establish sharp bounds on the costs and benefits of advances in data-processing technologies. We find that data-processing improvements have meaningfully increased total factor productivity (TFP) over the past two decades (3.8-4.4%) by enabling firms to better optimize their scale and product choice. Yet, the corresponding welfare gains have been modest (0.3-1.4%), as improvements in information also affect firms' information production and pricing strategies in ways that limit the welfare benefits. Our findings thus highlight an important role for corrective data regulation in the modern information economy.

To motivate our analysis, we use micro-level data on managerial forecasts from the I/B/E/S-Compustat panel. Using the merged data set, we document a systematic rise in the accuracy of U.S. firms' expectations over time. Over the past two decades, firms' expectations of one-year-ahead revenue have witnessed accuracy increases between 24-41%, with similar developments across most sectors and for other outcome variables (e.g., profits and capital expenditures). Crucially, these gains persist even after controlling for the volatility of firm-level and aggregate shocks, implying that firms have become substantially more informed since the early 2000s.

To better understand these developments, we examine the cross-sectional determinants of firms' accuracy. Across a range of specifications and controls, we find robust evidence that larger firms are, all else equal, more accurate than smaller ones, with substantial heterogeneity across the size distribution.[4] Indeed, using a simple simulation exercise, we show that most of the aggregate improvement in accuracy over the past two decades can be accounted for by changes along this size-accuracy relationship. In contrast, changes in sectoral composition and in the volatility of outcome variables play only a minor role.

We conduct a range of robustness checks which show that these results hold across alternative measures, specifications, and sets of firm-level controls. In addition, we document a positive relationship between accuracy of expectations and data expenditures using recent estimates from IDC's Data Wallet Tracker (IDC, 2021), which surveys firms' spending on hardware, software, and cloud computing. This evidence further supports the interpretation that the observed increase in accuracy reflects improvements in firms' information and data-processing capabilities, rather than changes in the underlying predictability of outcomes.

To explore the aggregate implications of our findings, we develop a macroeconomic model of information production. We consider an economy populated by heterogeneous consumers and firms. Consumers have preferences over differentiated varieties supplied by monopolistic

[4]See also Senga (2018), Tanaka *et al.* (2020), Chen *et al.* (2023), and Chen *et al.* (2024).

firms à la Dixit and Stiglitz (1977). The central friction in our economy is that firms must make production and pricing decisions under uncertainty about both *productivity and demand*. In particular, firms face uncertainty along three key margins. First, they are uncertain about their optimal *scale of operations*, as production depends on input choices as well as an ex-ante unknown productivity state. Second, they are uncertain about their optimal *product choice*, as firms seek to tailor their offerings to match unknown demand conditions. Third, demand uncertainty also implies that firms face uncertainty about their optimal *pricing strategies*.

A key feature of our framework is that firms can mitigate this uncertainty through costly information production. By using scarce resources, firms can obtain informative signals about the states of nature governing demand and productivity. We model advances in data processing as stemming either from a reduction in the cost of obtaining such signals or from an increase in their accuracy. This stylized approach captures key technological developments, such as declining computing and storage costs and improvements in processing speeds.[5]

An important component of our analysis is to construct two polar environments that span the spectrum of firms' ability to extract consumer surplus. In the first—the *baseline economy*—firms are exogenously constrained from price discrimination and their use of information is efficient. Information production facilitates the optimal determination of *scale* and *product choice*. In general equilibrium, it thus enhances aggregate TFP both directly—by improving firm-level efficiency—and indirectly—by reallocating inputs toward better-informed firms. In this case, advances in data processing unambiguously improve social welfare.

In the second—the *rent-extracting economy*—firms additionally use information to optimize their *pricing strategies*. In particular, we assume firms have full flexibility to use consumer information to segment consumers and tailor prices. We show that this ability alters firms' incentives to acquire information, distorts consumption allocations, and leads to an inefficient level of information production from a social perspective. Consequently, advances in data-processing technologies have ambiguous effects on both TFP and welfare in this environment.

We characterize the optimal corrective policy in the rent-extracting economy and show that, across a broad set of parameter values, firms engage in *excessive* information production from a social perspective. These findings lend support to policies such as the EUs General Data Protection Regulation and the Digital Services Act, which limit firms' ability to exploit consumer data in ways that facilitate discriminatory pricing. At the same time, we show that restricting data use may not always be desirable, and that policy instruments that are more targeted to certain types of information can be more effective.

Our framework highlights that the economy-wide effects of advances in data-processing

[5]See, e.g., Nordhaus (2008), Coyle and Hampton (2024), and Gill *et al.* (2024) for evidence of substantial declines in computing and storage costs and improvements in processing speeds over the past two decades.

technologies depend on key structural features of the environment. To assess the impact of improvements in data-processing technologies on the U.S. economy over the past two decades, we discipline the model using U.S. firm-level data and quantify their aggregate implications.

We proceed in two steps. First, we show that, consistent with the *qualitative* predictions of our model, in the data, more informed firms display higher and less dispersed (revenue-based measures of) total factor productivity, less dispersed marginal revenue products of labor and capital, and grow faster and larger. Second, we show that the *quantitative* relationships (i) between firm size and forecast accuracy, and (ii) between forecast accuracy and information production and spending in the model closely match their counterparts in the data.

Crucially, our quantitative results show that advances in data-processing technologies have led to sizable improvements in productive efficiency, raising aggregate TFP by 3.8–4.4% over the past two decades alone. Advances in data-processing have through this lens been an important contributor to growth over the sample period. That said, our results also suggest that the corresponding welfare gains have been more modest, between 0.3–1.4%. These bounds reflect the two benchmark environments described above. In the baseline economy, welfare gains are larger as advances enhance productive efficiency without introducing distortions. By contrast, in the rent-extracting economy, firms use information to inefficiently segment and price discriminate consumers, limiting welfare gains. In this environment, our estimates further suggest that firms allocate *excessive* resources to information production, as firms' incentives are driven by surplus extraction rather social efficiency. Our quantitative findings thus underscore a central role for data regulation in ensuring that advances in data-processing translate into welfare improvements. All in all, these two benchmarks represent polar extremes in firms' ability to extract consumer surplus. While the U.S. economy likely lies between these bounds, our framework provides a first step toward quantifying the macroeconomic consequences of advances in data-processing technologies.

Finally, we conduct two exercises that further assess the robustness of our quantitative results. First, we extend the framework to incorporate capital and variety accumulation—features that, in principle, should amplify the aggregate effects of advances in data processing. While these extensions modestly strengthen the effects on TFP, the welfare gains remain comparatively subdued. Second, we recalibrate the parameters governing the distribution of firms to match the end of our sample, allowing greater scope for changes in firm composition to account for TFP dynamics. Despite this, in both the baseline and the rent-extracting environments, our results remain unchanged, underscoring the robustness of our findings.

Our work builds on the growing macroeconomic literature on the rise of the data economy (e.g., Begenau *et al.*, 2018; Farboodi and Veldkamp, 2020, 2024; Eeckhout and Veldkamp, 2025; see also Baley and Veldkamp, 2025). Relative to this literature, we develop a quantitative

framework where information affects firms through three margins that have become especially salient in the data economy: scale, product choice, and pricing strategies. This structure allows us to decompose the macroeconomic effects of advances in data-processing technologies across these channels and study their welfare implications. Our paper is particularly closely related to David *et al.* (2016) and David and Venkateswaran (2019), who develop a methodology to measure firms' information use. Their empirical approach infers information use primarily from stock prices, whereas we use data on managers' expectations and IT expenditures to measure firms' information more directly. On the theory side, their analysis further focuses on the implications of information for factor misallocation, akin to our scale channel, while we emphasize that information also shapes firms' product choices and pricing decisions.

Our work, as such, also relates to the literature on price discrimination. Classic treatments are synthesized in Varian (1989), while Fudenberg and Villas-Boas (2012) surveys modern developments. Recent work, including Kehoe *et al.* (2018), Eeckhout and Veldkamp (2025), Farboodi *et al.* (2025), and Asriyan *et al.* (2025), studies how advances in data technologies enable firms to segment consumers and tailor prices. Relative to this literature, our contribution is to tractably embed discriminatory pricing into a general-equilibrium macroeconomic environment with endogenous information acquisition.[6] In our framework, information improves productive efficiency by helping firms align production decisions with fundamentals, but it also enables surplus extraction and can distort consumption allocations through discriminatory pricing.[7] In general equilibrium, discriminatory pricing changes firms' incentives to acquire information and drives a wedge between the private and the social value of data.

Finally, our work contributes to the macroeconomic literature on imperfect information, going back to Lucas (1972); see, among others, Woodford (2002), Mankiw and Reis (2002), Angeletos and Pavan (2007), Ordonez (2009), Lorenzoni (2009), Maćkowiak and Wiederholt (2009), and Angeletos *et al.* (2021). This literature studies how information frictions shape aggregate outcomes, typically taking the information structure as given. In contrast, we study firms' endogenous information choices in an environment where information affects scale, product choice, and pricing decisions. We show how such strategic information choices amplify the aggregate consequences of imperfect information and provide, to our knowledge, the first quantitative decomposition of how recent advances in data-processing technologies—and the associated decline in information frictions—affect macroeconomic outcomes.

[6] Our approach is also related to recent contributions studying the macroeconomic implications of non-linear pricing; see, e.g., Bornstein and Peter (2024) and Lorenzini and Martner (2026).

[7] This mechanism is related to the broader insight from the literature on trade under asymmetric information that additional (even public) information need not improve welfare in the presence of pre-existing distortions; see, for example, Malherbe (2012), Gorton and Ordonez (2014), and Asriyan *et al.* (2017).

2 Accuracy and Firm Information

We present new evidence on the accuracy of firm expectations over time and across the firm-size distribution. To start, we use micro data on firm expectations from the I/B/E/S managerial guidance database. The I/B/E/S data set contains, for an individual firm-year, a manager's publicly stated expectation of their firm's revenue, profits, and other performance variables for the upcoming year. We exploit one-year-ahead forecasts made concurrently with the release of the previous year's financials. We link the I/B/E/S database to Compustat, which provides detailed data on firms' financials. The merged I/B/E/S-Compustat sample spans the period 2002-2022. Appendix A.1 provides more information on the sample construction.

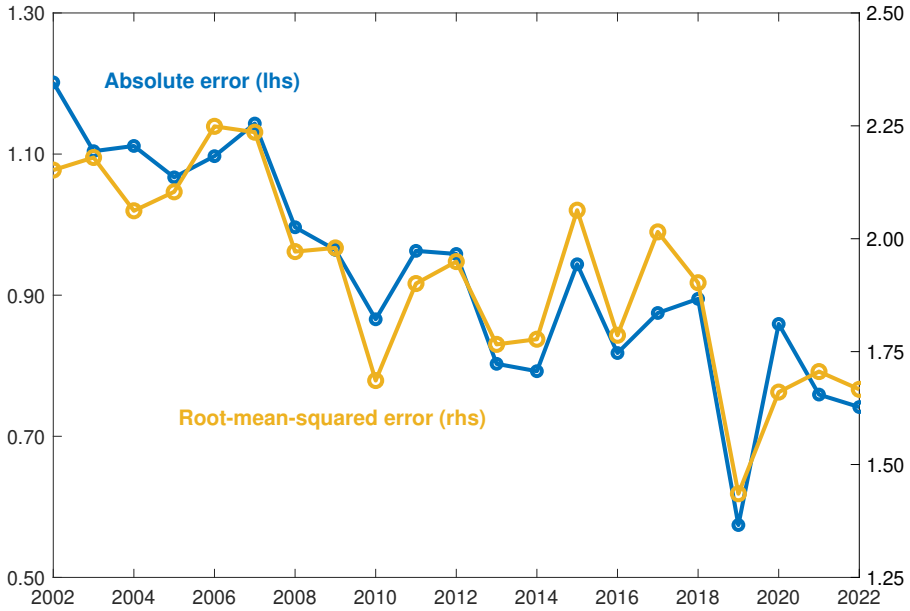
We begin by documenting changes in the accuracy of firms' expectations over time.[8] We focus on one-year-ahead revenue errors, defined as the realization minus its predicted value. A negative error thus corresponds to an over-estimate of future revenue. Figure 1 depicts the evolution of the average accuracy of firms' expectations over time. All else equal, over the past two decades, firms' expectations have improved markedly, with an average improvement in one-year-ahead accuracy between 24-41%, depending on the accuracy measure used. Firms' expectations have, on average, become substantially more accurate over time. Table A.3 in the Appendix confirms this initial finding using regressions of individual errors on time.

A natural candidate explanation for firms' increased accuracy is a decline in economy-wide volatility, such as that which occurred during the Great Moderation (e.g., Arias *et al.*, 2007). However, as Table A.4 in the Appendix shows, similar results hold after partialling out *sector*×*time* fixed effects from firms' errors. Consistent with most of the uncertainty faced by firms being due to *firm-specific* rather than *aggregate* factors (Lucas, 1977), the lion's share of accuracy increases here arises from improvements in firms' expectations about firm-specific outcomes. Combined with the observation that idiosyncratic volatility has, if anything, increased over our sample period (e.g., Bloom *et al.*, 2018 and Section 6), we conclude that *firms must have become substantially more informed since the early 2000s.*

The increased informativeness of firms is suggestive of changes in firm behavior or characteristics. The cross-section of firms can, as such, be revealing about the drivers behind the overall improvement over time. Because of a sizable fixed-cost component to the processing of information (e.g., Brynjolfsson *et al.*, 2008; Bloom *et al.*, 2019), it is natural to ask whether there is a relationship between firm size and the accuracy of firms' expectations. To investigate this question, Panel (a) in Figure 2 plots the difference between the average accuracy of one-

[8] In Appendix A.4, we conduct several data-validation tests, akin to those in Tanaka *et al.* (2020), Chen *et al.* (2023), and Chen *et al.* (2024). In particular, we show that firms' expectations are (close to) unbiased, feature a symmetric error distribution, that more (less) optimistic firms increase (decrease) their use of factors of production, and that positive (negative) surprises result in more (less) inputs being employed subsequently.

Figure 1: Time Evolution of Revenue Accuracy



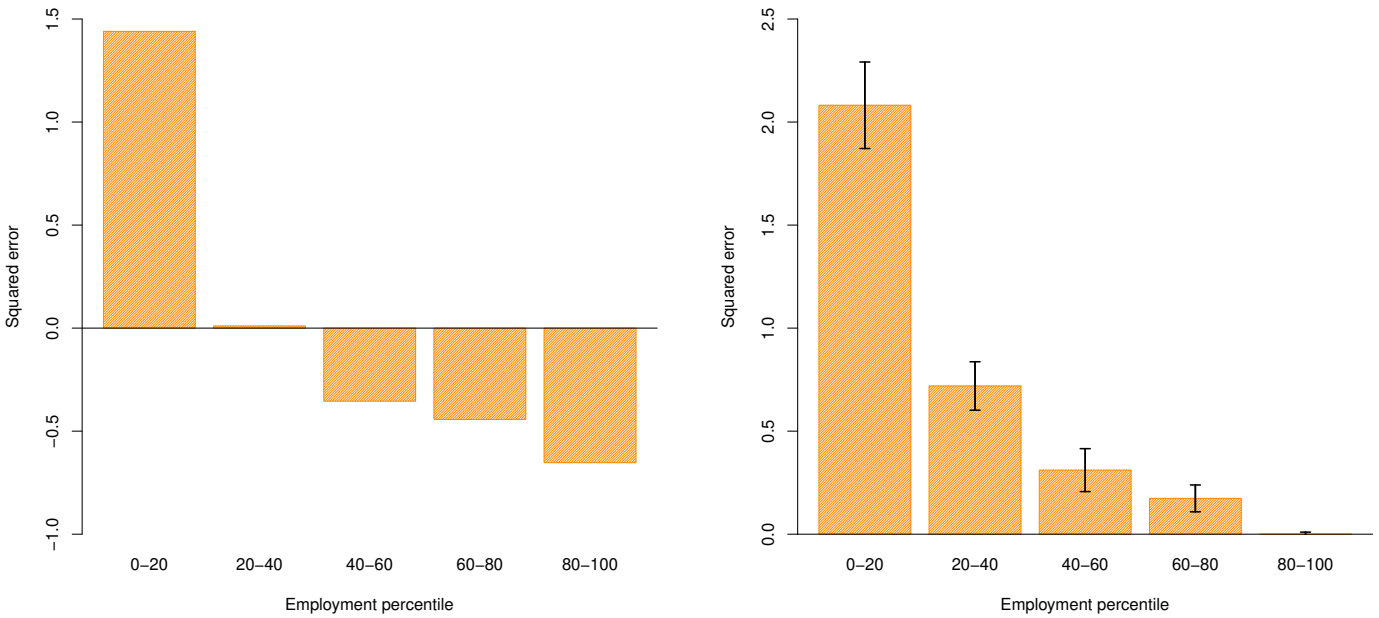
Note: Data from the I/B/E/S-Compustat sample. The panel shows the mean absolute error of one-year-ahead revenue forecasts on the left vertical axis, and the root mean-squared-error on the right axis. Revenue errors are scaled by a firm's tangible capital stock. Table A.3 in the Appendix shows the associated regression results.

year-ahead revenue expectations within size (employment) quintiles and the overall average taken across all firm sizes. The results show a marked, monotone relationship between firm size and the accuracy of firm expectations in the raw data. Larger firms have more accurate expectations—with an especially pronounced difference when moving away from the bottom quintile of the firm size distribution.

The relationship in Panel (a) may, nevertheless, be contaminated by other factors, such as differences in the volatility of the outcome variable across firm size or learning with age, which may be correlated with firm size for other reasons. To address this concern, Panel (b) in Figure 2 plots estimates from a regression of the accuracy of firm expectations on firm size, controlling for firm characteristics and time and sector fixed effects. Table I explores the effects of alternative controls and estimation methods, crucially, controlling for changes in the volatility of firm revenue and productivity over time. Our results confirm the findings from the raw data. The accuracy of expectations improves with size, even after controlling for firm characteristics. *Larger firms are, all else equal, more informed than smaller ones.*

Table A.5 in the Appendix shows that the documented size-accuracy relationship further extends to alternative data sets—in this case, the Duke-Richmond Fed CFO survey (Graham *et al.*, 2023; Appendix A.2)—which surveys firms' expectations of macroeconomic outcomes over which firms have no control. The latter is important as it further bolsters the case

Figure 2: Revenue Expectations Across the Size Distribution



Panel (a): Relative Accuracy

Panel (b): Coefficient on Size

Note: Panel (a) plots the difference between the squared error of one-year-ahead log-revenue expectations from the I/B/E/S-Compustat merger within size (employment) quintiles and the overall average taken across all size levels. Panel (b) plots the coefficient estimates on size from a regression of the squared value of individual errors on the size quintile the firm belongs to, controlling for firm characteristics (Table I Column 3). Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Whisker-intervals are one-standard deviation robust (clustered) confidence bounds. Sample: 2002-2022.

that accuracy improves with firm size due to improvements in firms' information rather than differences in the outcome variable across the size distribution.

The magnitude of the estimated effect of size in Table I is, moreover, considerable. Increasing the size of a firm by one quintile, for example, decreases the associated squared error by 47% of its average value (Column 1 in Table I). As documented in Figure A.2 and Tables A.3-A.4 in the Appendix, over the past two decades, firm size has increased drastically—with, for example, a close to doubling in the share of firms with employment exceeding the 80th percentile of the 2002-employment distribution.[9] Crucially, after controlling for this evolution in the firm-size distribution, the effect of time itself becomes insignificant (Column 2 in Table I). This suggest that accuracy increases unrelated to firm size have played only a secondary role over this period. Indeed, Figure A.4 in the Appendix, using estimates from Table I, shows that the change in the firm-size distribution alone can account for approximately 80% of the observed increase in overall accuracy. Clearly, these estimates cannot be interpreted as causal,

[9]See also, e.g., Autor *et al.* (2020), Kwon *et al.* (2023), among others, and Appendix Figure A.3.

Table I: Revenue Expectations, Firm Size, and Time

	Squared (log.) revenue errors					
	(1)	(2)	(3)	(4)	(5)	(6)
Firm size	-0.468^{***}	-0.454^{***}		-0.416^{***}	-0.290**	-0.430^{*}
	(0.055)	(0.052)		(0.122)	(0.124)	(0.226)
Firm size (1)			2.082^{***}			
			(0.210)			
Firm size (2)			0.719^{***}			
			(0.118)			
Firm size (3)			0.311^{***}			
			(0.104)			
Firm size (4)			0.174^{***}			
			(0.065)			
Time		0.007				
		(0.007)				
Firm age		-0.063^{**}	-0.072^{**}	0.118^{**}	0.187^{**}	-0.117
		(0.032)	(0.033)	(0.057)	(0.066)	(0.093)
Log rev. volatility				-0.030		
				(0.027)		
Log TFP. volatility					1.095	
					(0.744)	
Observations	12,489	12,489	12,489	6,809	5,637	2,570
Sector FE	✓	✓	✓	×	×	×
Firm FE	×	×	×	✓	✓	✓
Time FE	×	×	✓	✓	✓	✓
Panel GMM	×	×	×	×	×	✓
F statistic	3.911^{***}	3.922^{***}	4.322^{***}	9.295^{***}	8.005^{***}	NA.

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates of the squared value of one-year-ahead log-revenue errors on firm size (employment) and sector (NAICS-4) fixed effects. Firm size is measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Column (2) adds time and age controls. Column (3) allows for separate coefficients on size-levels (estimates are relative to the largest firms, those in the 80-100th percentile) and time fixed effects. Column (4) allows for firm fixed effects instead of sector fixed effects and controls for the individual four-year-rolling revenue volatility. Column (5) instead controls for individual four-year-rolling TFP volatility (Appendix A.2). Finally, Column (6) provides Arellano-Bond estimates. Revenue errors are scaled by firm capital and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022. *p<0.1,**p<0.05,***p<0.01

as firm size and the accuracy of expectations are likely determined jointly. Yet, our results demonstrate that, *over the past two decades, an intimate relationship has existed between the accuracy of firms' expectations, firms' information, and firms' size.* Such an intimate bidirectional relationship will be a key feature of our model framework in Section 3.

We obtain similar estimates to those in Figures 1 and 2 for alternative measures of size and accuracy. Table A.6 and Figure A.5 show similar estimates when proxying size with firm assets instead of employment. Tables A.7-A.8 document that an alternative measure of accuracy (e.g., the absolute error) likewise monotonically increases with firm size, irrespective of whether size is measured by employment or assets. We further find that across all-but-one sector accuracy has improved over time and increases with size (Figure A.6).

Tables A.9-A.13 in the Appendix contain additional analysis. We document that our findings extend to firm expectations of other variables than revenue (profits and capex), and to different transformations of firm revenue. We also show that the accuracy of firm expectations improves after large acquisitions of other firms, lending support to the notion that there are increasing returns to information; and extends to different assumptions about sectoral and time fixed effects. Finally, although direct measures of firms' overall data expenditures are not available for the U.S. economy, we provide suggestive evidence linking firms' data-related investments to their accuracy. Table A.13 shows that firms with higher expenditures on hardware, software, and cloud computing (IaaS) exhibit more accurate expectations in 2021. These estimates are obtained by merging the IDC IT Wallet Database—which provides firm-level estimates of IT expenditures using a combination of top-down industry allocations and bottom-up firm data (Appendix A.3)—with the I/B/E/S-Compustat sample. Consistent with this pattern, Table A.12 further shows that firms with larger stocks of *acquired intangible capital* in I/B/E/S-Compustat (Chiavari and Goraya, 2023)—which includes business software and data-processing system expenditures—also exhibit more accurate expectations. Combined, these results provide direct support for the mechanism emphasized below: firms that invest more in data infrastructure and processing exhibit more accurate expectations.

In summary, the evidence in this section documents that firms' accuracy has improved substantially over the past two decades and is systematically higher for larger firms. Over the same period, firms that invest more heavily in data-processing technologies also exhibit more accurate expectations. Taken together, our results point to an important role for firms' information acquisition strategies. In what follows, we develop a model in which firms invest in information about demand and productivity as part of their broader production and pricing decisions. This framework allows us to then study how information acquisition interacts with firm scale, product choice, and pricing in general equilibrium.

3 The Baseline Economy

We start by developing our *baseline economy*, a central feature of which is that firms are uncertain about the optimal *scale* of their operations and their optimal *product choice*. We study the setting in which information also affects firms' *pricing strategy* in Section 5.

3.1 Environment

We consider an economy populated by a mass of households, indexed by $i \in [0,1]$, with CES-utility preferences over a continuum of differentiated varieties, indexed by $j \in [0,1]$:

$$\mathcal{U}_i \;=\; C_i \;\equiv\; \left(\int_0^1 \left(\delta_{ij} \cdot c_{ij}\right)^{\frac{\theta-1}{\theta}}\, dj\right)^{\frac{\theta}{\theta-1}}, \tag{1}$$

where $c_{ij} \geq 0$ is household i's consumption of variety j, $\delta_{ij} > 0$ is the household's demand shifter, and $\theta > 1$ is the elasticity of substitution across varieties. Each household is endowed with $N > 0$ units of labor, which is the economy's only factor of production.

Demand Shifters. The demand shifter δ_{ij} is distributed identically across households, and consists of a common, variety-level component and a household–variety–specific component:

$$\log \delta_{ij} = \log \sigma_j + \log \varsigma_{ij}. \tag{2}$$

The common component σ_j takes values in $\{\bar{\sigma}, \underline{\sigma}\}$ with $\bar{\sigma} > \underline{\sigma} > 0$, depending on the match between the *type of variety* j available in the market, denoted by $x_j \in \{\text{red}, \text{blue}\}$, and a random *demand state* $\omega_j \in \{\text{red}, \text{blue}\}$, where $\mathbb{P}(\omega_j = \text{red}) = \frac{1}{2}$. We assume variety j commands high demand when its type matches the demand state, and low demand otherwise:

$$\sigma_j = \bar{\sigma} \quad \text{if and only if} \quad x_j = \omega_j.$$

The component ς_{ij}, by contrast, captures *household taste heterogeneity* for a given variety. It is independently distributed across households and varieties, and independent of ω_j. We further assume ς_{ij} takes values in $\{h, \ell\}$, with $h > \ell > 0$, where $\gamma \equiv \mathbb{P}(\varsigma_{ij} = h) \in (0,1)$.[10]

Production Technology. Each variety is produced by a monopolistically competitive firm, owned by households. Firm j chooses the type x_j of its variety at no cost and produces the quantity y_j of it using labor in accordance with the linear production technology:

$$y_j = A_j \cdot n_j, \tag{3}$$

[10] Appendix B provides a simple microfoundation for this demand structure by modeling household preferences over red and blue variety types explicitly.

where n_j are the units of labor employed by the firm and A_j is the firm's productivity. A firm's productivity is in turn comprised of two components:

$$\log A_j = \mu_j + \upsilon_j, \tag{4}$$

where $\mu_j \sim \mathcal{N}(0, \tau_\mu^{-1})$ and $\upsilon_j \sim \mathcal{N}(0, \tau_a^{-1})$ are independent across firms and of each other. The parameters τ_μ^{-1} and τ_a^{-1} govern the cross-sectional dispersion of each component.[11]

Uncertainty and Information. The central friction in the economy is that firms are imperfectly informed about both *productivity* and *demand* when making production decisions. When choosing x_j and n_j, firm j observes its mean-productivity level μ_j but not the productivity innovation υ_j, the demand state ω_j, nor the realizations of household-specific taste shocks $\{\varsigma_{ij}\}_i$. As a result, the component μ_j is a source of *ex ante* heterogeneity across firms, while the shocks υ_j, ω_j, and $\{\varsigma_{ij}\}_i$ are sources of *ex post* uncertainty.[12]

A novel feature of our framework is that a firm can produce information to overcome its uncertainty about productivity and demand. In particular, firm j can obtain signals:

$$s_j^\upsilon \;=\; \upsilon_j + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}\Big(0,\, [\tau_j^\upsilon]^{-1}\Big), \tag{5}$$

$$s_j^\omega \in \{\text{red}, \text{blue}\}, \qquad \mathbb{P}(s_j^\omega = \omega_j \mid \omega_j) = \tau_j^\omega \in \left[\tfrac{1}{2}, 1\right], \tag{6}$$

and

$$s_{ij}^\varsigma \in \{h, \ell\}, \qquad \mathbb{P}(s_{ij}^\varsigma = \varsigma_{ij} \mid \varsigma_{ij}) = \tau_{ij}^\varsigma \in \left[\tfrac{1}{2}, 1\right], \quad \forall i. \tag{7}$$

where all signal errors are independent across signals and across firms. We assume that a firm can obtain signals with baseline precisions $(\underline{\tau}^\upsilon, \underline{\tau}^\omega, \underline{\tau}^\varsigma)$ at no cost; however, by allocating $\chi > 0$ units of labor to information production, the firm can increase the precision of its signals to $(\bar{\tau}^\upsilon, \bar{\tau}^\omega, \bar{\tau}^\varsigma) > (\underline{\tau}^\upsilon, \underline{\tau}^\omega, \underline{\tau}^\varsigma)$. We denote firm j's information choice by $\iota_j \in \{0, 1\}$, where $\iota_j = 1$ indicates information production. To save on notation, in what follows, we let:

$$\boldsymbol{s}_j \equiv \Big(s_j^\upsilon, s_j^\omega, \{s_{ij}^\varsigma\}_i\Big), \qquad \boldsymbol{\tau}_j \equiv \Big(\tau_j^\upsilon, \tau_j^\omega, \{\tau_{ij}^\varsigma\}_i\Big).$$

3.2 Timing and Markets

Timing. The economy unfolds over three stages. In *Stage 1*, each firm j observes its mean productivity level μ_j and chooses whether or not to produce information, $\iota_j \in \{0, 1\}$. The

[11] We note that there is an isomorphic formulation of our economy where—rather than scaling firm j's productivity—the shock υ_j scales the preferences of all consumers of variety j, i.e., $\delta_{ij} = \exp^{\upsilon_j} \cdot \sigma_j \cdot \varsigma_{ij}$.

[12] Ex-ante heterogeneity is not essential for our results. It is, however, useful to "purify" potential mixed-strategy equilibria, and to better connect the theory to firm-level data. See Section 2 for further discussion.

economy then proceeds to *Stage 2*, where, conditional on its information choice, each firm observes its signals $\boldsymbol{s}_j$ and chooses the amount of labor to employ, n_j, as well as the type of variety to produce, x_j. The economy concludes in *Stage 3*. At this stage, each household i learns its tastes $\{\delta_{ij}\}_j$ and supplies labor, while each firm j learns the productivity shock υ_j, produces its chosen variety, and trades with all households as described next.

Markets. The labor market is perfectly competitive: households supply labor and firms hire labor, taking the wage w as given. In the market for variety j, the monopolistic firm is uncertain about consumers' preferences and observes signals about demand. In our baseline framework, we adopt a pricing protocol that is standard in macroeconomic models with imperfect information: given the quantity produced, each variety trades at a *uniform* unit price p_j that clears the market given the realized demand (see, e.g., Angeletos *et al.*, 2016).

3.3 Remarks on Modeling Approach

In our baseline framework, information affects firm behavior through two channels.

First, the production of information helps a firm learn its productivity state, υ_j, and hence better choose its overall *scale* of production. This channel captures the canonical approach to modeling information frictions in macroeconomics, going back to at least Lucas (1972).

Second, the production of information also helps a firm learn its demand state, ω_j, and hence allocate its factors of production towards the variety-type most preferred by consumers on average. Although we have modeled this channel through *product choice*, it is isomorphic to several natural alternatives: e.g., the allocation of factors between different plants, or input sourcing from different suppliers.[13] Common to all these interpretations is that the production of information helps a firm improve its internal factor allocation.

Combined, the two channels through which information affects firm behavior are thought to be among the most important ways by which advances in data-processing technologies have improved firm decision-making over the past two decades (Brynjolfsson and McElheran, 2016; Ali *et al.*, 2020; Adams *et al.*, 2025; Veldkamp and Chung, 2024; Abis and Veldkamp, 2024). Yet, a common concern among academics and policymakers is that advances in data-processing technologies may also have facilitated discriminatory practices vis-à-vis consumers. We capture this third channel of information in Section 5, where the production of information about consumer-specific demand $\{\varsigma_{ij}\}$ also allows firms to optimize their *pricing strategies* so as to better extract rents from consumers, albeit potentially at a social cost. As we will show,

[13] See, for example, the different business cases reported in O'Neill (2023). See also The Wall Street Journal's report on Levi's, whose use of information technologies to analyze fashion trends among young consumers was instrumental in the introduction of 'baggy (Gen-Z) jeans' (https://www.wsj.com/articles/how-tech-helped-levis-ride-the-baggy-jeans-trend-f290721d).

the interaction between all three channels shapes the normative properties of our economy and yields valuable lessons about the desirability of data regulation.

Finally, our baseline economy features both a fixed supply of production factors (i.e., labor) and a fixed set of firms or varieties. In Section 6.3, we show that standard approaches to introduce factor supply elasticity through capital accumulation and through entry and exit of varieties merely amplify the aggregate consequences of information production.

3.4 Optimization and Equilibrium

Household Problem. Household $i \in [0,1]$ chooses consumption of individual varieties $\{c_{ij}\}_{j\in[0,1]}$ in *Stage 3* to maximize its utility in (1) subject to the budget constraint:

$$\int_0^1 p_j \cdot c_{ij} \cdot dj = w \cdot N + \int_0^1 \pi_j \cdot dj, \tag{8}$$

where π_j are the profits from firm j. The household takes prices, $\{p_j\}_{j\in[0,1]}$, and the wage rate, w, as given when solving its problem. The solution yields:

$$c_{ij} = \delta_{ij}^{\theta-1} \cdot p_j^{-\theta} \cdot C_i, \tag{9}$$

where we have normalized the ideal price index to one. Since all households are ex-ante identical, $C_i = C$ for all i, and the aggregate demand for variety j becomes:

$$c_j \equiv \int_0^1 c_{ij} \cdot di = \delta_j^{\theta-1} \cdot p_j^{-\theta} \cdot C \quad \text{with} \quad \delta_j \equiv \delta(\sigma_j) = \left(\gamma \cdot h^{\theta-1} + (1-\gamma) \cdot \ell^{\theta-1}\right)^{\frac{1}{\theta-1}} \cdot \sigma_j. \tag{10}$$

Thus, it is *as if*, for each variety j, there is a *representative consumer* with demand shifter δ_j; but where this shifter depends on whether the firm produces the preferred type ($\sigma_j \in \{\bar{\sigma}, \underline{\sigma}\}$).

Firm Problem. The ex-post profits of firm $j \in [0,1]$ are given by the firm's revenue net of its expenditures on labor and information:

$$\pi_j = p_j \cdot y_j - w \cdot n_j - w \cdot \chi \cdot \iota_j. \tag{11}$$

In *Stage 2*, conditional on its information set $(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)$, the firm chooses labor, n_j, and variety-type, x_j, to maximize its expected profits from goods production:

$$\widehat{\pi}_j\left(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right) \equiv \max_{n_j, x_j} \mathbb{E}\left[p_j \cdot y_j - w \cdot n_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right], \tag{12}$$

subject to feasibility in (3) and the output of the variety being equal its demand in (10). Here, $\mathbb{E}\left[\cdot|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]$ denotes the expectations operator conditional on $(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)$. When choosing

n_j, the firm optimally equates its expected marginal revenue product of labor to the wage:

$$\frac{\theta-1}{\theta}\cdot\frac{\mathbb{E}\left[p_j\cdot y_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]}{n_j}=w. \tag{13}$$

When choosing the variety-type, the firm in turn optimally sets:

$$x_j=\arg\max_{x_j\in\{\text{red,blue}\}}\mathbb{E}[p_j\cdot y_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j]. \tag{14}$$

In *Stage 1*, conditional on its mean-productivity, μ_j, the firm chooses information production, ι_j, so as to maximize its expected profits. The firm produces information when the expected increase in profits (from goods production) exceeds the cost of information:

$$\iota_j\begin{cases}=1 & \text{if } \mathbb{E}\left[\widehat{\pi}_j\left(\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right)|\mu_j,\boldsymbol{\tau}_j=\bar{\boldsymbol{\tau}}\right]-w\cdot\chi\geq\mathbb{E}\left[\widehat{\pi}_j\left(\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right)|\mu_j,\boldsymbol{\tau}_j=\underline{\boldsymbol{\tau}}\right]\\ =0 & \text{otherwise}\end{cases}, \tag{15}$$

where $\mathbb{E}\left[\cdot|\mu_j,\boldsymbol{\tau}_j\right]$ denotes the expectations operator conditional on μ_j and $\boldsymbol{\tau}_j$.

Equilibrium Notion. An equilibrium consists of allocations $\left\{\left\{c_{ij}\right\}_{i\in[0,1]},y_j,x_j,n_j,\iota_j\right\}_{j\in[0,1]}$, prices $\left\{p_j\right\}_{j\in[0,1]}$, and a wage w such that:

- Given the prices and the wage, the allocations solve the household and the firm problems, i.e., Equations (8)-(15) hold, and;
- Given the allocations, the goods and the labor markets clear, i.e., $c_j=y_j$ for all $j\in[0,1]$ and $\int_0^1(n_j+\chi\cdot\iota_j)\cdot dj=N$.

4 Equilibrium Characterization

We proceed by analyzing firm-level outcomes in partial equilibrium, taking all economy-wide variables as given (Section 4.1). We then characterize the aggregate consequences of firms' information choices in general equilibrium (Section 4.2). We conclude by analyzing the aggregate effects of improvements in data processing within our baseline framework (Section 4.3).

4.1 Firm-Level Implications of Information

A noticeable advantage of our framework is that all firm-level choices are driven by two simple objects. The first captures the relevant notion of *market size* faced by firms in our economy and encodes all relevant general-equilibrium interactions. The second, by contrast, measures the profitability boost that a firm can expect if it were to produce information.

Definition 1. *Let* $\Omega \equiv C \cdot \left(\frac{\theta}{\theta-1} \cdot w\right)^{-\theta}$ *be the* ***market size*** *faced by firms in the economy.*

Market size summarizes the overall demand conditions faced by firms. It increases with aggregate consumption, C, and decreases with the wage, w, which determines marginal production costs. Market size is a key determinant of firms' information production choices in conjunction with what we refer to as firms' *information shifter*.

Definition 2. *For* $\boldsymbol{\tau} = (\tau^{\upsilon}, \tau^{\omega}, \tau^{\varsigma})$*, define firms'* ***information shifter*** *as:*

$$g(\boldsymbol{\tau}) \equiv \left[\tau^{\omega} \cdot \delta(\bar{\sigma})^{\frac{\theta-1}{\theta}} + (1-\tau^{\omega}) \cdot \delta(\underline{\sigma})^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \cdot \exp^{\frac{1}{2}\cdot\frac{\theta-1}{\theta}\cdot\frac{1}{\tau_a}\cdot\frac{\tau_a+\theta\cdot\tau^{\upsilon}}{\tau_a+\tau^{\upsilon}}}, \tag{16}$$

where $\delta(\cdot)$ *is given by Equation (10).*

As we will see, $g(\cdot)$ summarizes all the benefits that information has both at the firm level and in the aggregate. We now characterize firms' optimal production and information choices.

Proposition 1. *In any equilibrium:*

(i) Firm j *with mean productivity* μ_j*, which observes signals* $\boldsymbol{s}_j$ *with precisions* $\boldsymbol{\tau}_j$*, chooses:*

$$x_j = s_j^{\omega} \quad \text{and} \quad n_j = \mathbb{E}\left[\left(\delta_j \cdot A_j\right)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot \Omega. \tag{17}$$

(ii) Firm j *with mean productivity* μ_j *chooses:*

$$\iota_j = \begin{cases} 1 & \text{if} \quad \frac{1}{\theta-1} \cdot \left(\mathbb{E}\left[n_j|\mu_j, \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}\right] - \mathbb{E}\left[n_j|\mu_j, \boldsymbol{\tau}_j = \underline{\boldsymbol{\tau}}\right]\right) \geq \chi \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

where:

$$\mathbb{E}\left[n_j|\mu_j, \boldsymbol{\tau}_j\right] = \exp^{(\theta-1)\cdot\mu_j} \cdot g(\boldsymbol{\tau}_j)^{\theta-1} \cdot \Omega. \tag{19}$$

Recall that information affects firms along two margins. First, it allows firms to better match their product type to consumer demand; and, second, it allows firms to adjust production scale to realized productivity. Proposition 1 formalizes how these two channels shape firm behavior. It furthermore has two sets of important implications.

1) Information and Resource Allocation. Part (i) of Proposition 1 implies that information boosts the efficiency of resource allocation both *within* and *across* firms.

First, by helping a firm allocate its factors of production towards the variety-type most preferred by households, information increases firm-level total-factor productivity. To see this, observe that for given input and information choices (x_j, n_j, ι_j), the total surplus (or utility)

generated by firm j rises monotonically with $(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}}$.[14] Thus, an appropriate measure of *(log-)total factor productivity* for firm j is:

$$\text{tfp}_j \equiv \frac{\theta-1}{\theta} \cdot \log(\delta_j \cdot A_j). \tag{20}$$

Because the demand shifter faced by firm j, δ_j, is affected by the firm's choice of product type, x_j, the distribution of tfp_j becomes endogenous to the firm's information choice:

Corollary 1. *For all j, the conditional mean and variance of tfp is:*

$$\mathbb{E}[\text{tfp}_j|\mu_j, \boldsymbol{\tau}_j] = \frac{\theta-1}{\theta} \cdot \left(\tau_j^{\omega} \cdot \log\left[\delta(\bar{\sigma})\right] + (1-\tau_j^{\omega}) \cdot \log\left[\delta(\underline{\sigma})\right] + \mu_j\right) \tag{21}$$

$$\mathbb{VAR}[\text{tfp}_j|\mu_j, \boldsymbol{\tau}_j] = \left(\frac{\theta-1}{\theta}\right)^2 \cdot \left[\tau_j^{\omega} \cdot (1-\tau_j^{\omega}) \cdot \left(\log\left[\delta(\bar{\sigma})\right] - \log\left[\delta(\underline{\sigma})\right]\right)^2 + \frac{1}{\tau_a}\right], \tag{22}$$

where $\delta(\cdot)$ is given by Equation (10).

All else equal, firms with more precise demand information, i.e., a larger τ_j^{ω}, exhibit both higher mean and lower volatility of tfp. This is because firms with more precise demand information are able to tailor their products more effectively to consumer tastes, thereby raising and stabilizing their effective demand. We note that a variant of this "internal factor allocation" channel of information also appears in Farboodi and Veldkamp (2024), albeit modeled in a reduced-form manner that directly modifies the process for firm-level tfp.

Second, by helping a firm correlate its overall employment of factors with the realized shocks to the firm's tfp, information also reduces (ex-post) factor misallocation across firms. To see this, let mrp_j denote the *(log-)marginal-revenue product* of labor for firm j:

$$\text{mrp}_j \equiv \log(p_j \cdot y_j) - \log(n_j). \tag{23}$$

Absent information frictions, $\text{mrp}_j = \text{mrp}_{j'}$ for all j, j' (Equation (13)). Any dispersion in mrp_j is thus a tell-tale sign of inefficiency resulting from information frictions.

Corollary 2. *For all j, the conditional variance of mrp_j is:*

$$\begin{aligned}\mathbb{VAR}[\text{mrp}_j|\mu_j, \boldsymbol{\tau}_j] &= \mathbb{VAR}[\text{tfp}_j|\mu_j, \boldsymbol{\tau}_j] + \left(\frac{\theta-1}{\theta}\right)^2 \cdot \left(\frac{1}{\tau_a + \tau_j^{v}} - \frac{1}{\tau_a}\right) \\ &= \mathbb{VAR}[\text{error}_j|\mu_j, \boldsymbol{\tau}_j],\end{aligned} \tag{24}$$

[14] The total surplus generated by firm j is equal to the total utility generated by the firm, $(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}$, net of its labor costs, $w \cdot (n_j + \chi \cdot \iota_j)$, all measured in the numeraire.

where $\text{error}_j \equiv \log(p_j \cdot y_j) - \log\left(\mathbb{E}[p_j \cdot y_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j]\right)$.

All else equal, factor misallocation is therefore lower among firms with more precise information, i.e., a larger τ_j^v and τ_j^ω. This is because firms with more precise information are also those that have more accurate revenue expectations. Indeed, the cross-sectional dispersion in mrp_j is equal to the variance of a firm's log-revenue error, a mapping we will later utilize to pin down the implied decline in mrp-dispersion from the observed increase in firms' accuracy in the data. We note that the role of information through its effect on factor misallocation has also been studied in David *et al.* (2016) and David and Venkateswaran (2019), albeit in a setting where firms do not themselves decide on their information production.

Lastly, notice that in the baseline economy, a firm's information about households' idiosyncratic taste components $\{\varsigma_{ij}\}_i$ does not affect its production decisions nor has implications for resource allocation. Accordingly, the information shifter $g(\boldsymbol{\tau})$, which summarizes a firm's benefit from information production, is independent of τ^ς. This feature will change once we allow firms to also use information to optimize their pricing strategies (Section 5).

2) Information and Firm-size Distribution. Part (ii) of Proposition 1 implies that—by improving efficiency of resource allocation—information also alters the *firm-size distribution*.

To see this, consider the benefits to a firm from producing information, which are given by the change in the firm's expected profits from goods production resulting from a more efficient resource allocation (Corollaries 1 and 2). From the optimality condition (13), a firm's expected profits from goods production are proportional to its expected employment:

$$\mathbb{E}\left[p_j \cdot y_j - w \cdot n_j | \mu_j, \boldsymbol{\tau}_j\right] = \frac{1}{\theta - 1} \cdot w \cdot \mathbb{E}[n_j | \mu_j, \boldsymbol{\tau}_j]. \tag{25}$$

It then follows from Equation (19) in Proposition 1 that the information shifter $g(\cdot)$ is key to understanding the benefits of information. In particular, given the properties of $g(\cdot)$:

$$\frac{\partial \mathbb{E}[n_j | \mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^v} > 0, \ \frac{\partial \mathbb{E}[n_j | \mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^\omega} > 0, \ \frac{\partial^2 \mathbb{E}[n_j | \mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^v \partial \tau_j^\omega} > 0. \tag{26}$$

Thus, not only does each channel of information alone—whether learning about demand or productivity—boost a firm's size and profits, but the two channels interact and *reinforce* each other, amplifying the overall size gains from information production.[15]

We have so far shown that, holding fixed a firm's ex-ante mean productivity, μ_j, information production—by improving the efficiency of resource allocation—boosts that firm's expected size and profitability. In equilibrium, however, there is also a selection of firms into information

[15] In fact, *all three channels* of information interact and reinforce each other once we also incorporate the pricing channel in Section 5.

production, which we must consider when studying any relationship between information and firm characteristics. Since the benefits from information production scale with a firm's expected size—which grows with μ_j (Equation (19))—but the information cost χ does not, it is the ex-ante more productive firms that choose to produce information. As it turns out, such a selection, if anything, reinforces the effects of information production outlined above.

Corollary 3. *In equilibrium, firm j produces information if and only if:*

$$\mu_j \geq \bar{\mu} \equiv \frac{1}{\theta - 1} \cdot \log \left[\frac{(\theta - 1) \cdot \chi}{\left(g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \right) \cdot \Omega} \right]. \tag{27}$$

Further, more informed firms, i.e., $\{j : \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}\}$, have on average higher and less dispersed tfp, less dispersed mrp, and they grow larger and more profitable as measured by:

$$\mathbb{E}[\mathcal{Z}_j | \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}] > \mathbb{E}[\mathcal{Z}_j | \boldsymbol{\tau}_j = \underline{\boldsymbol{\tau}}], \tag{28}$$

for $\mathcal{Z}_j \in \{n_j, n_j + \chi \cdot \iota_j, p_j \cdot y_j, p_j \cdot y_j - w \cdot n_j, \pi_j\}$.

To conclude, it is worth noting that, although in our setting selection into information production is driven by ex-ante productivity differences, alternative specifications are also possible. For example, the cross-sectional patterns described in Corollary 3 would also arise if we instead assumed that firms were heterogeneous in information costs (i.e., in χ_j); or if there were no heterogeneity whatsoever but we instead focused on the parameter region in which the equilibrium is in mixed strategies (i.e., where an interior share of firms produce information). Compared to these alternatives, our current approach, nevertheless, has the clear advantage of allowing us to more closely discipline the link between firm size and the accuracy of firms' information, using our empirical findings documented in Section 2.

We have characterized how information production affects firm choices and outcomes at the micro level. The next natural step is to study the general equilibrium of the economy. We return to validate the cross-sectional predictions, described in Corollaries 1-3 above, using the I/B/E/S-Compustat sample in Section 6.

4.2 Aggregate Implications of Information

We begin by defining an appropriate notion of aggregate *total factor productivity* (TFP), which captures the economy's allocative efficiency given firms' information sets—namely, the amount of consumption (or utility) the economy generates from labor allocated to goods production.

Definition 3. *Given aggregate consumption, C, and aggregate employment in goods production, $\mathcal{N} \equiv \int_0^1 n_j \cdot dj$, we define aggregate* ***total factor productivity*** *as $\mathcal{A} \equiv C \cdot \mathcal{N}^{-1}$.*

Using households' goods demands, market clearing for variety j, and firms' labor choices in Proposition 1, we can express TFP solely as a function of firms' information sets, $\{(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)\}_j$, and hence their information choices:

$$\mathcal{A} = \left(\int_0^1 (\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot \left(\frac{n_j}{\mathcal{N}} \right)^{\frac{\theta-1}{\theta}} \cdot dj \right)^{\frac{\theta}{\theta-1}} = \left(\int_0^1 \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j \right]^{\theta} \cdot dj \right)^{\frac{1}{\theta-1}}, \tag{29}$$

where to obtain the last equality we have made use of the fact that:

$$\mathcal{N} = \int_0^1 \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j \right]^{\theta} \cdot dj \cdot \Omega. \tag{30}$$

Since from Corollary 3, we know that firms' information choices are fully pinned down by the mean productivity $\bar{\mu}$ of the marginal firm that is just indifferent to producing information (henceforth, the *marginal-type*), we have that:

Lemma 1. *Given the marginal-type, $\bar{\mu}$, aggregate total factor productivity, $\mathcal{A}$, equals:*

$$\mathcal{A}(\bar{\mu}; g) \equiv \exp^{\frac{\theta-1}{2} \cdot \frac{1}{\tau_\mu}} \cdot \left[g(\underline{\boldsymbol{\tau}})^{\theta-1} \cdot (1 - \xi(\bar{\mu})) + g(\bar{\boldsymbol{\tau}})^{\theta-1} \cdot \xi(\bar{\mu}) \right]^{\frac{1}{\theta-1}}, \tag{31}$$

where $\xi(\bar{\mu}) \equiv \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu} + (\theta - 1) \cdot \frac{1}{\sqrt{\tau_\mu}} \right)$ and $\Phi(\cdot)$ is the standard normal c.d.f.

The main implication of Lemma 1 is that, by boosting the efficiency of resource allocation—both within and across firms—information production raises the productivity of the economy as a whole. Importantly, the weight $\xi(\bar{\mu})$ in Equation (31) captures the share of information producing firms in the economy, adjusted for the fact that these firms are larger (Corollary 3) and thus have a larger impact on the aggregate allocation of resources. As more firms produce information—that is, as $\bar{\mu}$ falls and $\xi(\bar{\mu})$ rises—more firms feature the larger information shifter and, as a result, aggregate TFP rises (since $g(\bar{\boldsymbol{\tau}}) > g(\underline{\boldsymbol{\tau}})$ by Definition 2).[16]

Aggregate productivity is an important determinant of market size, $\Omega = C \cdot \left(\frac{\theta}{\theta-1} \cdot w \right)^{-\theta}$, and hence of all economy-wide objects in our economy. Combining the definition of TFP with Equation (30) and the labor market-clearing condition shows that:

$$\Omega = \frac{\mathcal{A}(\bar{\mu}, g) \cdot \left[N - \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu} \right) \cdot \chi \right]}{\mathcal{A}(\bar{\mu}, g)^{\theta}}. \tag{32}$$

Recall from Section 4.1 that a firm's incentive to produce information depends on the market size, Ω. Equation (32) shows that, in general equilibrium, the market size itself also depends on the collective information choices of firms, as summarized by the marginal-type, $\bar{\mu}$.

[16] Notice that we have made the dependence of TFP on the information shifter, $g(\cdot)$, explicit in Equation (31), a notational device that will later prove useful in Section 5.

Information production affects market size through its effect both on consumption and on firms' production costs. Beginning with costs, the denominator in Equation (32) reflects the fact that greater information production (i.e., a lower $\bar{\mu}$) raises firms' marginal costs. This occurs because higher TFP—driven by information—increases labor demand and thus the equilibrium wage ($w = \frac{\theta-1}{\theta} \cdot \mathcal{A}$). Turning to consumption, the numerator in Equation (32) captures two opposing forces. On the one hand, more information directly raises aggregate consumption by boosting the economy's TFP. On the other hand, the final term in the numerator reflects a countervailing effect: information production diverts scarce labor away from goods production, thereby depressing consumption. It is straightforward to see that, since $\theta > 1$, the net effect of these opposing forces is negative, and market size decreases with information production (i.e., as $\bar{\mu}$ falls). This implies that, in equilibrium, firms' information production choices are *strategic substitutes*, guaranteeing the uniqueness of the equilibrium.

A convenient feature of our framework is that the economy's equilibrium can be studied through the intersection of just two schedules. Equation (32) defines a continuous schedule $\Omega : \mathbb{R} \to \mathbb{R}^+$, which maps a given marginal-type $\bar{\mu}$ into a market size $\Omega(\bar{\mu})$ that is consistent with market clearing. By contrast, Equation (27) defines a continuous schedule $\bar{\mu} : \mathbb{R}^+ \to \mathbb{R}$ that for a given market size, Ω, yields the marginal-type $\bar{\mu}(\Omega)$ that is indifferent to producing information. An equilibrium is given by the fixed point of the composite map: $\bar{\mu} \circ \Omega : \mathbb{R} \to \mathbb{R}$. The following proposition states that such a fixed point exists and is unique.

Proposition 2. *An equilibrium exists, is unique, and in it the marginal-type that is just indifferent to producing information solves:*

$$\bar{\mu}(\Omega(\mu^\star)) = \mu^\star, \tag{33}$$

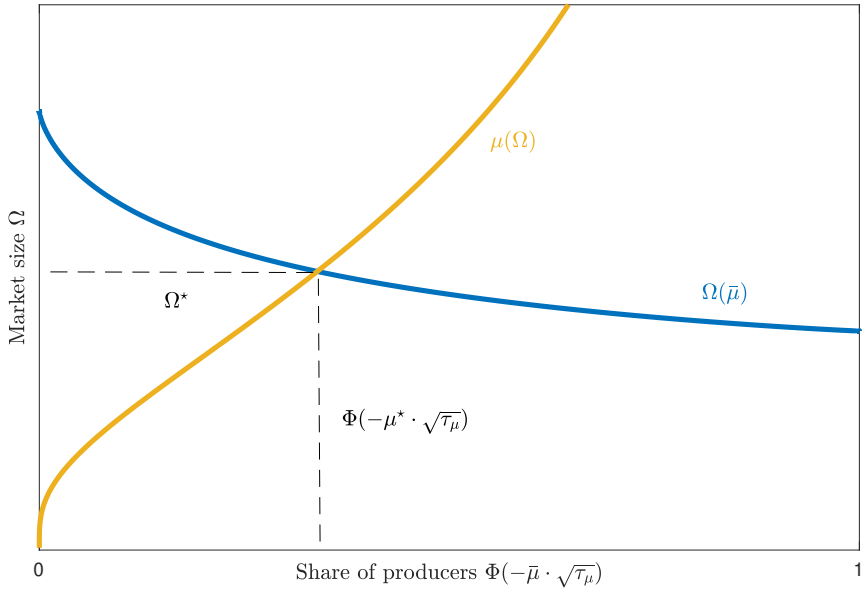
where $\bar{\mu}(\cdot)$ and $\Omega(\cdot)$ are defined by Equations (27) and (32), respectively. Aggregate TFP and consumption in turn equal:

$$\mathcal{A}^\star = \mathcal{A}(\mu^\star, g) \quad \text{and} \quad C^\star = \mathcal{A}^\star \cdot \left[N - \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right) \cdot \chi\right], \tag{34}$$

where $\mathcal{A}(\cdot)$ is stated in Lemma 1 and $\Phi(\cdot)$ is the standard normal c.d.f.

The equilibrium's determination is illustrated in Figure 3, which depicts the relationship between market size, Ω, and the share of information producers, as given by $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$. The orange locus depicts the combinations of $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$ and Ω that are consistent with the indifference condition in Equation (27). This locus is upward sloping because, as we discussed in Section 4.1, a firm's incentive to produce information increases with the market size. The blue locus, by contrast, depicts the combinations of $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$ and Ω, which are consistent

Figure 3: Equilibrium Determination



Note: The upward-sloping (orange) locus depicts the relationship between market size, Ω, and share of firms that produce information, $\Phi(-\bar{\mu} \cdot \sqrt{\tau_\mu})$, as defined by Equation (27). The downward-sloping (blue) locus instead depicts the relationship between Ω and $\Phi(-\bar{\mu} \cdot \sqrt{\tau_\mu})$ as defined by Equation (32).

with market clearing, as described in Equation (32). As we discussed above, this locus is instead downward sloping. An equilibrium is characterized by the unique intersection of the two, which pins down the equilibrium pair $(\mu^\star, \Omega^\star)$.

4.3 Aggregate Effects of Advances in Data Processing

We now use our baseline framework to examine the macroeconomic consequences of advances in data-processing technologies. The past two decades have seen firms increasingly adopt and rely on large-scale, data-intensive algorithms to enhance their economic decision-making (e.g., Baley and Veldkamp, 2025). This shift in the nature by which firms make their economic choices has, in turn, been driven by substantial declines in computing costs and improvements in processing speeds (e.g., Nordhaus, 2008; Coyle and Hampton, 2024; Gill *et al.*, 2024).

We analyze the aggregate effects of such technological advances through two comparative static exercises: (i) a reduction in the information cost parameter, χ; and (ii) an increase in the precision of information that can be obtained through information production, $\bar{\tau}^{\upsilon}$ and $\bar{\tau}^{\omega}$.[17] These changes serve as stylized representations of declines in computing costs and

[17] In the baseline economy, the precision of information about consumers' idiosyncratic-taste component τ^{ς} is also irrelevant for aggregate TFP and consumption.

improvements in processing capabilities. The following proposition summarizes their effects:

Proposition 3. *An improvement in data-processing technologies, such as a fall in χ, a rise in $\bar{\tau}^{\upsilon}$, or a rise in $\bar{\tau}^{\omega}$, leads to an increase in the share of firms producing information, $\Phi\left(-\mu^{\star}\cdot\sqrt{\tau_{\mu}}\right)$, an increase in aggregate TFP, $\mathcal{A}^{\star}$, and consumption, $C^{\star}$.*

Figure 4 illustrates the equilibrium effects of improvements in data-processing technologies on the share of information-producing firms, $\Phi(-\mu^{\star}\cdot\sqrt{\tau_{\mu}})$, and on the market size, $\Omega^{\star}$, respectively. Panel (a) depicts the impact of a decline in χ, while Panel (b) illustrates the consequences of an increase in $\bar{\tau}^{\upsilon}$ or $\bar{\tau}^{\omega}$.[18] Both improvements make information production more attractive from a firm's perspective (Proposition 1), leading to a rightward shift in the orange loci in Figure 4. However, their general equilibrium consequences differ.

A reduction in χ shifts the blue locus upward, as lower information costs allow the economy to allocate more labor to goods production, thereby increasing consumption and expanding, all else equal, market size, $\Omega(\mu^{\star})$. In contrast, an increase in $\bar{\tau}^{\upsilon}$ or $\bar{\tau}^{\omega}$ shifts the blue locus downward, as the higher accuracy of firms' information enhances TFP, which, as discussed in Section 4.2, has a depressing effect on the equilibrium market size, $\Omega(\mu^{\star})$.

Despite these differing mechanisms, the overall macroeconomic effects of the two technological improvements are nevertheless similar. In both cases, aggregate TFP and consumption increase. This is not surprising given the benign nature of the information assumed so far. Indeed, as we formally establish in Proposition 6 below the equilibrium of our baseline economy is efficient. As a result, within this framework, improvements in data-processing technologies are inherently beneficial. As we show next, however, these efficiency properties change markedly once information is also allowed to influence firms' pricing behavior.

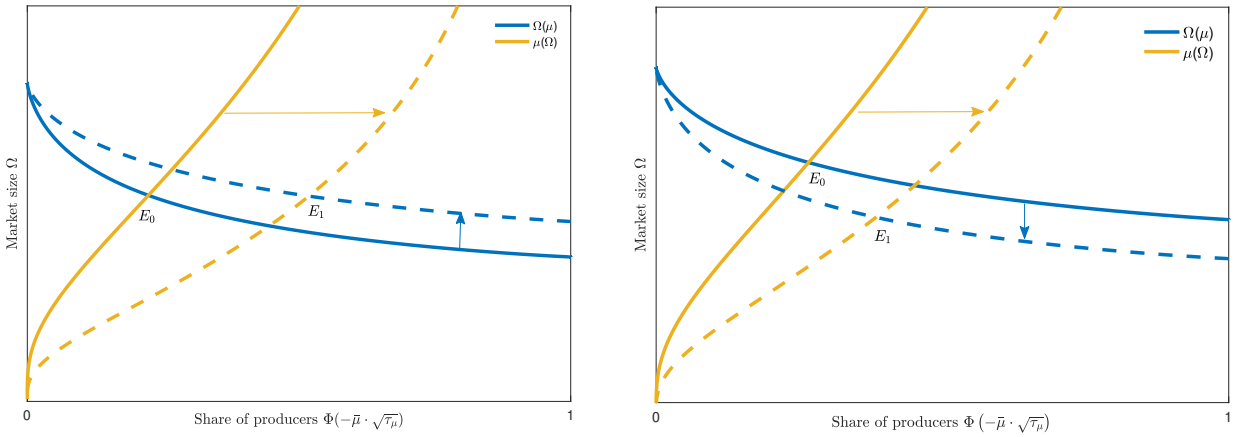
5 The Social Value of Information and Pricing

Thus far, we have emphasized the benign role of information in improving resource allocation within and across firms. Information production, however, need not be universally beneficial. A growing concern surrounding recent advances in data-processing technologies is that they may enhance firms' ability to engage in increasingly sophisticated *price discrimination* (OECD, 2018; European Commission, 2020).[19] To address this concern, we extend our framework to allow information to influence firms' pricing decisions. We refer to this extended framework as the *rent-extracting economy*. A key normative distinction relative to the baseline is that, while information continues to improve firms' production and product-choice decisions, privately

[18] The aggregate effects of an increase in $\bar{\tau}^{\upsilon}$ are similar to those of an increase in $\bar{\tau}^{\omega}$.

[19] See, for example, O'Neill (2023) and Eeckhout and Veldkamp (2025).

Figure 4: Improvements in Data-Processing Technologies



Panel (a): a fall in χ Panel (b): a rise in $\bar{\tau}^{\upsilon}$ or $\bar{\tau}^{\sigma}$

Note: Panel (a) depicts the effects of a fall in the cost of information, χ, whereas Panel (b) depicts the effects of a rise increase in the accuracy of information, $\bar{\tau}^{\upsilon}$ and $\bar{\tau}^{\sigma}$. The solid (dashed) loci depicts the relationship between market size, Ω, and share of firms producing information, $\Phi(-\bar{\mu}\cdot\sqrt{\tau_{\mu}})$, before (after) the improvement in information technologies. Equilibria before (after) the change are denoted by E_0 (E_1).

optimal pricing responses may now distort allocations. Improvements in data processing can, as a result, have ambiguous effects on welfare, creating scope for data regulation.

5.1 Pricing with Information

In the baseline economy, firms trade with consumers at *uniform prices*, a pricing protocol that has become canonical in macroeconomics. By contrast, in the rent-extracting economy we allow firms to condition their trading arrangements on *all* available information and to implement the *optimal trading mechanism*. This benchmark captures an environment in which firms fully exploit their information to maximize surplus extraction. Together, the two economies span polar benchmark environments: the baseline restricts firms to uniform pricing independent of consumer information, whereas the rent-extracting economy allows firms to fully tailor trading arrangements using available data. In what follows, we describe the main features of the firm's problem and relegate detailed derivations to Appendix C.1. Apart from the pricing protocol, the environment is otherwise identical to that in the baseline economy.

We make two changes to our baseline setup. First, after observing signals $\{s_{ij}\}_i$ about consumer-specific idiosyncratic tastes (or "types" ς_{ij}), we assume that firm j assigns consumers to *segments* $s \in \{h, \ell\}$ as a function of the realized signals.[20] In the limiting case of perfectly

[20] To economize on notation we drop superscript ς from the signals $\{s^{\varsigma}_{ij}\}_i$.

informative signals, the induced segmentation coincides with the underlying consumer types. Second, we assume that the firm offers a *menu* of allocations $\mathcal{M}_j(s) = \{(t_j(\varsigma|s), q_j(\varsigma|s))\}_{\varsigma\in\{h,\ell\}}$ to consumers in segment $s \in \{h, \ell\}$, who then decide whether and which allocation in the menu to accept. A consumer who selects allocation $(t_j(\varsigma|s), q_j(\varsigma|s))$ receives $q_j(\varsigma|s)$ units of variety j in exchange for a payment of $t_j(\varsigma|s)$; otherwise, no trade occurs. Thus, while segmentation is based on realized signals, allocations within each segment are determined by consumers' self-selection, rather than by directly conditioning on consumer types.

A growing literature shows that firms use consumer data—such as purchase histories, past browsing, and location and device characteristics—to segment consumers and tailor prices and offers accordingly. Audit studies further document systematic price and offer variation across consumer groups in online markets, while industry surveys indicate that a substantial share of firms adopting AI for personalization already use these tools for real-time pricing and promotions (Hannak *et al.*, 2014; OECD, 2018; Adams *et al.*, 2025). The above modifications allow us to tractably study such price discrimination. While uniform pricing has become standard in macroeconomics, there is no widely accepted framework for incorporating discriminatory pricing into general equilibrium models. Our approach—which combines consumer-specific information with optimal mechanism design—provides a tractable benchmark that embeds information-dependent price discrimination into an otherwise canonical CES environment.

Importantly, within this extension, information about consumer-specific types is valuable to a firm *only insofar as* it enables segmentation into groups with systematically different demand characteristics. Absent segmentation, such information has no effect on a firm's choices or its ability to extract surplus, as in the baseline economy. Moreover, under our CES specification with divisible goods—preferences that are canonical in macroeconomics—segmentation alone does not affect equilibrium outcomes if firms are restricted to uniform pricing, since all consumers share the same price elasticity of demand. Segmentation enables additional surplus extraction only if firms can implement non-linear (menu) pricing. Rather than depart from these canonical preferences, we below retain the CES structure and instead analyze price discrimination through optimal menu pricing.[21]

Crucially, our main insights do not hinge on this particular pricing environment. In Appendix C.2, we show that they also obtain in an alternative specification—commonly studied in the microeconomics literature on price discrimination—in which goods are indivisible and firms post uniform prices within each segment (e.g., Fudenberg and Villas-Boas, 2012). In this setting, firms can still segment consumers based on observable signals but, due to indivisibilities, must charge a single price within each segment. Nevertheless, we show that the

[21] Recent work by Bornstein and Peter (2024) documents the prevalence of non-linear pricing in consumer-pricing data and its implications for standard measures of allocative efficiency. Likewise, Lorenzini and Martner (2026) document and explore the implications of widespread non-linear pricing in firm-to-firm transactions.

resulting patterns of surplus extraction and allocative distortions are qualitatively similar to those in our main formulation.

5.2 Optimal Trading Mechanisms

Building on the above pricing environment, we now characterize a firm's optimal trading mechanism and the associated pricing rule in the rent-extracting economy.

At *Stage 3*, the firm chooses a mechanism to maximize revenues, as all costs are sunk. These revenues in turn reflect the payments collected from each consumer type in each segment:

$$\sum_{s=h,\ell} \mathbb{P}(s_{ij} = s) \cdot \sum_{\varsigma=h,\ell} \mathbb{P}(\varsigma_{ij} = \varsigma | s_{ij} = s) \cdot t_j(\varsigma|s), \tag{35}$$

where $(t_j(\varsigma|s), q_j(\varsigma|s))$ denotes the allocation selected by a type-$\varsigma \in \{h, \ell\}$ consumer assigned to segment $s \in \{h, \ell\}$, $\mathbb{P}(s_{ij} = s)$ is the share of consumers in that segment, and $\mathbb{P}(\varsigma_{ij} = \varsigma | s_{ij} = s)$ is the share of type-ς consumers within the segment. The firm's problem is subject to standard incentive compatibility, individual rationality, and feasibility constraints.

Since the firm does not perfectly observe individual consumer types, in equilibrium, each type in each segment must optimally select its intended allocation (*incentive compatibility*):

$$(t_j(\varsigma|s), q_j(\varsigma|s)) = \arg \max_{(t,q) \in \mathcal{M}_j(s)} (\sigma_j \cdot \varsigma \cdot q)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t \quad \text{for} \quad \varsigma, s \in \{h, \ell\}, \tag{36}$$

where the right-hand side is the consumer's surplus from accepting an allocation (t, q) in menu $\mathcal{M}_j(s)$.[22] To participate in the mechanism, each consumer type in each segment must further obtain a non-negative surplus from the allocation (*individual rationality*):

$$(\sigma_j \cdot \varsigma \cdot q_j(\varsigma|s))^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j(\varsigma|s) \geq 0 \quad \text{for} \quad \varsigma, s \in \{h, \ell\}. \tag{37}$$

Finally, as the firm supplies its variety inelastically at *Stage 3*, the aggregate quantity allocated to consumers cannot exceed the quantity that the firm has produced (*feasibility*):

$$\sum_{s=h,\ell} \mathbb{P}(s_{ij} = s) \cdot \sum_{\varsigma=h,\ell} \mathbb{P}(\varsigma_{ij} = \varsigma | s_{ij} = s) \cdot q_j(\varsigma|s) \leq A_j \cdot n_j. \tag{38}$$

The solution to the firm's problem takes a well-known form. At the optimum, the individual rationality constraint of ℓ-type consumers, the incentive-compatibility constraint of h-type

[22] Note, we assume that at the pricing stage (*Stage 3*) firm j observes the demand state ω_j, and hence the common component of demand, σ_j. This assumption entails no loss of generality: as is well known, the common-value component ω_j can be costlessly elicited by conditioning allocations on the full vector of reported types (e.g., Crémer and McLean, 1988).

consumers, and the feasibility constraint all bind. Given these constraints, quantities are chosen to equalize marginal revenues across consumer types and segments. The key trade-off is a familiar one: extracting additional surplus from high-demand consumers requires distorting the allocation offered to low-demand consumers. Information about consumers types, in turn, enhances the firm's rent-extraction ability by allowing it to better separate different types.

Let $\alpha_j(\varsigma|s) \equiv q_j(\varsigma|s)/(A_j \cdot n_j)$ denote the share of output allocated to a type-ς consumer in segment s. The optimal allocations are then given by:

$$\alpha_j(h|h) = \alpha_j(h|\ell) = \frac{h^{\theta-1}}{\gamma \cdot h^{\theta-1} + (1-\gamma) \cdot \ell^{\theta-1} \cdot \sum_{s=h,\ell} \cdot \mathbb{P}(s_{ij} = s|\varsigma_{ij} = \ell) \cdot \psi_j(s)^{\theta}}, \tag{39}$$

$$\alpha_j(\ell|s) = \alpha_j(h|s) \cdot \left(\frac{\ell}{h}\right)^{\theta-1} \cdot \psi_j(s)^{\theta} \text{ for } s \in \{h, \ell\}, \tag{40}$$

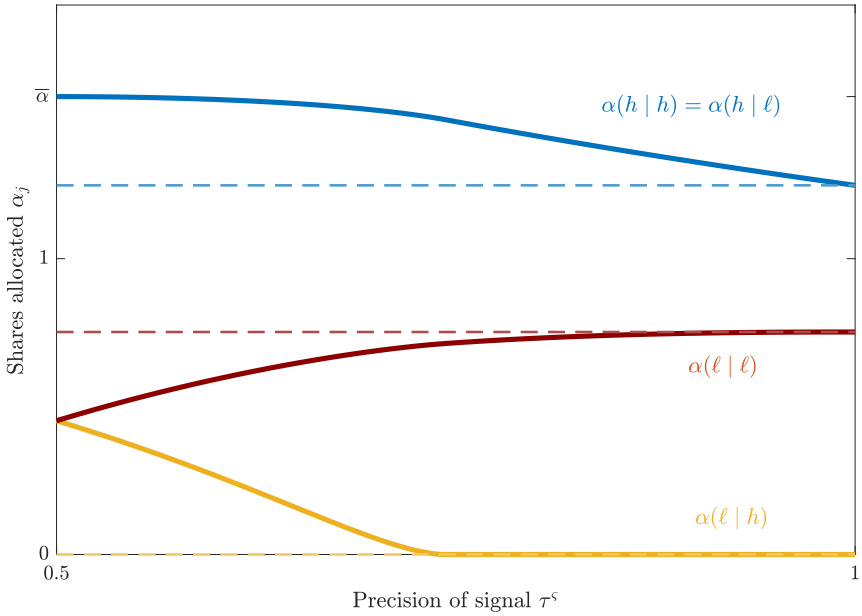
where:

$$\psi_j(s) \equiv \begin{cases} \frac{1-\mathbb{P}(\varsigma_{ij}=h|s_{ij}=s)\cdot\left(\frac{h}{\ell}\right)^{\frac{\theta-1}{\theta}}}{1-\mathbb{P}(\varsigma_{ij}=h|s_{ij}=s)} & \text{if } \ell^{\frac{\theta-1}{\theta}} \geq \mathbb{P}(\varsigma_{ij} = h|s_{ij} = s) \cdot h^{\frac{\theta-1}{\theta}} \\ 0 & \text{otherwise} \end{cases} \text{ for } s \in \{h, \ell\}. \tag{41}$$

The shares $\alpha_j(\varsigma|s)$ describe how output is distributed across consumer types and segments, while the term $\psi_j(s)$ provides the associated micro-level distortion. Indeed, $\psi_j(s)$ is the ratio of the marginal utility of consumption of h-type consumer to that of ℓ-type consumer in segment s. Allocative efficiency requires this ratio to equal one. Instead, $\psi_j(s) \leq 1$ reflects the distortion implied by a binding incentive-compatibility constraint. As Equation (41) shows, the magnitude of this distortion depends on the composition of consumer types—and is thus itself endogenously shaped by the precision of firms' information about consumer tastes, τ_j^{ς}.

When information is imprecise (i.e., $\tau_j^{\varsigma} = \frac{1}{2}$), the firm restricts trade with low-demand consumers to extract surplus from high-demand consumers. Because signal realizations are nevertheless uninformative about consumer types, the two segments are indistinguishable: thus, $\alpha_j(h|h) = \alpha_j(h|\ell) > \alpha_j(\ell|h) = \alpha_j(\ell|\ell)$ (Figure 5). As precision increases, the two segments become compositionally distinct—segment h contains relatively more high-demand consumers, segment ℓ relatively more low-demand ones. The firm responds asymmetrically: in segment h, where it now faces more high-demand consumers, it further restricts trade with low-demand types to extract more surplus from high-demand types, worsening allocative distortions. In segment ℓ, the opposite occurs—the firm is willing to trade more with low-demand types and extract less surplus from high-demand types, reducing allocative distortions. As a result, the net effect of information on total surplus is ambiguous, a point we come back to in the next subsection. In the limit as $\tau_j^{\varsigma} \to 1$, consumers become perfectly segmented, the

Figure 5: Allocative Distortions



Note: The figure depicts the output shares $\alpha_j(\varsigma|s)$ in Equation (39) as a function of the information precision, τ_j^{ς}, and the signal realization, $s = h$ or $s = \ell$. The dashed loci depict the efficient allocations, i.e., those that equalize the marginal utilities of consumption across consumer types.

firm extracts full surplus from each type, and allocations converge to efficiency as the measure of low-demand (high-demand) consumers in segment h (ℓ) vanishes.[23]

5.3 Rent Extraction and Firm-Level Inefficiencies

The upshot of the above compact characterization is that all firm-level choices can be summarized as functions of just two demand shifters. First, the ex-post profits of firm j—which govern the firm's optimal information and input choices in *Stages 1* and *2*—equal:

$$\pi_j = \left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot n_j - w \cdot \chi \cdot \iota_j, \tag{42}$$

so that it is *as if* firm j faces a representative consumer with *revenue-based demand shifter*:

$$\delta_j^R = \delta^R(\tau_j^{\varsigma}, \sigma_j) \equiv \left[\gamma \cdot h^{\theta-1} + (1-\gamma) \cdot \ell^{\theta-1} \cdot \sum_{s=h,\ell} \mathbb{P}(s_{ij} = s|\varsigma_{ij} = \ell) \cdot \psi_j(s)^{\theta}\right]^{\frac{1}{\theta-1}} \cdot \sigma_j. \tag{43}$$

[23] Throughout, the quantities allocated to high-demand consumers are identical across segments ($\alpha(h|h) = \alpha(h|\ell)$), reflecting the optimal equalization of their marginal utilities of consumption. Prices, however, differ across segments, with high-demand consumers facing a higher per-unit price in segment h than in segment ℓ. For sufficiently high precision, low-demand consumers are fully excluded from trade in segment h (Figure 5).

This demand shifter captures the portion of the total surplus that the firm extracts from consumers through its optimal pricing strategy.[24] Notice that it is fully pinned down by the common demand component σ_j and the precision of the firm's information τ_j^ς, as the latter determines the distribution $\{\mathbb{P}(s_{ij}|\varsigma_{ij})\}$ and hence the distortions $\{\psi_j(s_{ij})\}$.

Second, the ex-post total surplus—defined as profits plus consumer surplus—created by firm j can by contrast be expressed as:

$$u_j = \left(\delta_j^S \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot n_j - w \cdot \chi \cdot \iota_j, \tag{44}$$

where the *surplus-based demand shifter* that the firm faces is given by:

$$\delta_j^S = \delta^S(\tau_j^\varsigma, \sigma_j) \equiv \Psi(\tau_j^\varsigma) \cdot \delta^R(\tau_j^\varsigma, \sigma_j). \tag{45}$$

The wedge $\Psi_j \geq 1$ between the two demand shifters is pinned down by the precision of the firm's information about consumer tastes, τ_j^ς, and it concisely captures the misalignment between profit and surplus maximization that is inherent to the rent-extracting economy.[25]

The two demand shifters, δ_j^R and δ_j^S, and the wedge between them, Ψ_j, are central to the workings of the rent-extracting economy, where we note the analogy to the demand shifter δ_j in Equation (10) in our baseline framework. We summarize their key properties below:

Lemma 2. *Fix $\sigma_j > 0$. The revenue-based demand shifter, δ_j^R, increases monotonically with τ_j^ς and converges, as $\tau_j^\varsigma \to 1$, to the surplus-based demand shifter, $\delta_j^S \geq \delta_j^R$. Although δ_j^S reaches its maximum and Ψ_j its minimum at $\tau_j^\varsigma = 1$, both can be non-monotonic in τ_j^ς.*

Lemma 2 highlights a potential conflict between *allocative efficiency* and *surplus extraction*. The revenue-based shifter δ_j^R increases monotonically with information precision, as more precise segmentation always enables firms to extract surplus more effectively. In the limit, as $\tau_j^\varsigma \to 1$, segmentation becomes perfect and the firm captures the entire surplus, so that $\delta_j^R \to \delta_j^S$. By contrast, the surplus-based shifter δ_j^S may be non-monotonic in τ_j^ς, as improvements in information simultaneously affect surplus extraction and allocative distortions.

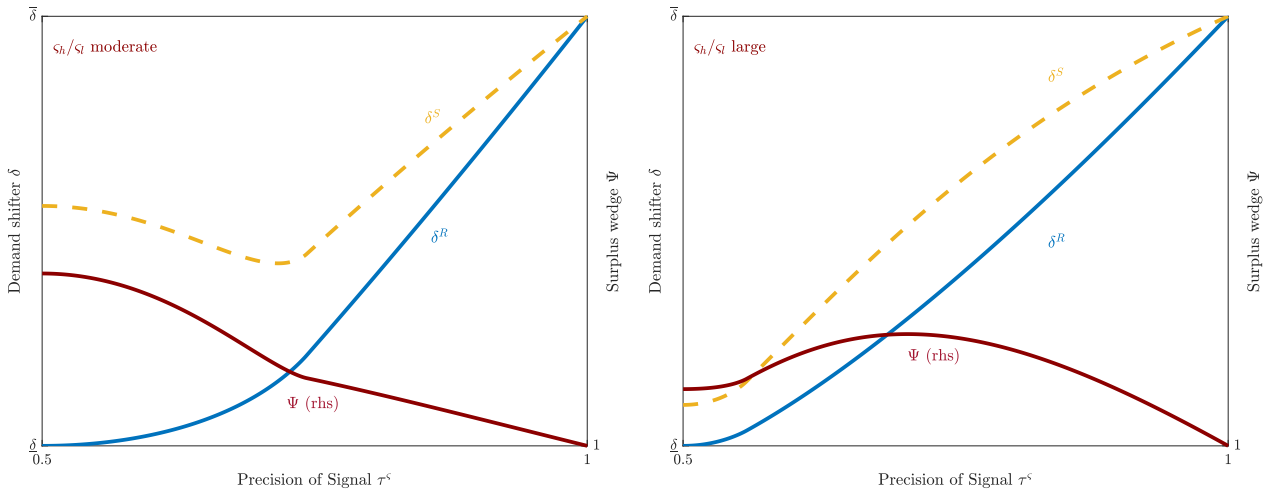
To see this, recall that an increase in τ_j^ς to start raises allocative distortions in segment h—by further restricting trade with low-demand consumers—while reducing distortions in

[24] The expression for δ_j^R follows by summing the payments $t_j(\varsigma|s)$ across consumer types ς and segments s.

[25] The expression for δ_j^S follows by summing the utilities $\left(\sigma_j \cdot \varsigma_j \cdot q_j\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}$ across consumer types ς and segments s, which implies that:

$$\Psi_j = \Psi(\tau_j^\varsigma) \equiv \left[1 + \frac{(1-\gamma) \cdot \ell^{\theta-1} \cdot \sum_{s=h,\ell} \mathbb{P}\left(s_{ij} = s|\varsigma_{ij} = \ell\right) \cdot \psi_j\left(s\right)^{\theta-1} \cdot \left(1 - \psi_j\left(s\right)\right)}{\gamma \cdot h^{\theta-1} + (1-\gamma) \cdot \ell^{\theta-1} \cdot \sum_{s=h,\ell} \mathbb{P}\left(s_{ij} = s|\varsigma_{ij} = \ell\right) \cdot \psi_j\left(s\right)^{\theta}}\right]^{\frac{\theta}{\theta-1}}.$$

Figure 6: Revenue- vs Surplus-Based Demand Shifters



(a) Moderate heterogeneity

(b) Large heterogeneity

Note: The figure depicts the demand shifters δ_j^S and δ_j^R, and the wedge Ψ_j as a function of τ_j^ς. Panel (a) plots the case in which h/ℓ is relatively small, while Panel (b) plots the case in which h/ℓ is relatively high.

segment ℓ by making trade with such consumers more efficient (Figure 5). When allocative distortions are initially small—e.g., when taste heterogeneity, captured by the ratio h/ℓ, is moderate—the first effect dominates (Panel (a) of Figure 6). In this case, more accurate information exacerbates distortions and reduces surplus, creating a conflict between efficiency and surplus extraction. By contrast, when distortions are initially large—e.g., when taste heterogeneity is large—the second effect dominates, and greater information precision both reduces distortions and raises total surplus (Panel (b)). In this case, no such conflict arises. Finally, Figure 6 also shows that the wedge $\Psi_j = \delta_j^S/\delta_j^R$ may vary non-monotonically with τ_j^ς.

When a firm produces information, the accuracy of its information about consumer-specific tastes increases from $\underline{\tau}^\varsigma$ to $\bar{\tau}^\varsigma$. To organize the results that follow, we will make use of Lemma 2 and partition the parameter space according to how this increase affects total surplus as well as the wedge between profit and surplus maximization.

Definition 4. *Segmentation is* ***socially destructive*** *(****valuable****) if δ_j^S decreases (increases) as τ_j^ς goes from $\underline{\tau}^\varsigma$ to $\bar{\tau}^\varsigma$. It is* ***severe*** *(****mild****) if Ψ_j decreases (increases) over the same range.*

We note that, because δ_j^R is increasing in τ_j^ς, Definition 4 implies that socially destructive segmentation is necessarily severe, whereas mild segmentation is necessarily socially valuable.

5.4 Distortions in General Equilibrium

We are now ready characterize the general equilibrium of the rent-extracting economy. Relative to the baseline framework, the only modification to the equilibrium conditions stems from

firms' use of *revenue-based* rather than *surplus-based* incentives when making their information and input choices. Indeed, let us define the modified, revenue-based information shifter:

$$g^R(\boldsymbol{\tau}) \equiv \left[\tau^\omega \cdot \delta^R(\tau^\varsigma, \bar{\sigma})^{\frac{\theta-1}{\theta}} + (1-\tau^\omega) \cdot \delta^R(\tau^\varsigma, \underline{\sigma})^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \cdot \exp^{\frac{1}{2}\cdot\frac{\theta-1}{\theta}\cdot\frac{1}{\tau_a}\cdot\frac{\tau_a+\theta\cdot\tau^\upsilon}{\tau_a+\tau^\upsilon}}. \tag{46}$$

This object is the analog of the information shifter $g(\cdot)$ in the baseline economy, with the sole difference that the demand shifter δ_j is replaced by the revenue-based demand shifter δ_j^R.

With this modification, the equilibrium characterization closely parallels that of the baseline economy. An equilibrium is defined in the same way as before, except exchange between firms and consumers at *Stage 3* occurs according to the trading mechanism in Section 5.1.

Proposition 4. *An equilibrium of the rent-extracting economy exists, is unique, and in it:*

(i) Firm j's choices of x_j, n_j and ι_j are given by Proposition 1, with the only modification that the demand shifter δ_j in Equation (17) is replaced by δ_j^R.

(ii) The equilibrium marginal-type, $\mu^\star$, is given by Proposition 2, with the only modification that the information shifter $g(\cdot)$ in Equations (27) and (32) is replaced by $g^R(\cdot)$.

Thus, the firms' equilibrium behavior in the rent-extracting economy is qualitatively similar to that in the baseline economy. This result is not surprising. As we have shown in Section 5.3, the only difference in the relevant revenue/profit expressions is that firms behave *as if* they face the revenue-based shifter δ_j^R instead of the baseline shifter δ_j.

An immediate implication of Proposition 4 is that the firm-level predictions summarized in Corollary 1-3 continue to hold. There is, however, one important caveat. When a conflict between allocative efficiency and surplus extraction arises, revenue-based measures of productivity and factor misallocation need not be informative about welfare. Although more informed firms earn higher revenues and profits, they may generate lower surplus. As a consequence, aggregate outcomes can differ sharply from those in the baseline economy.

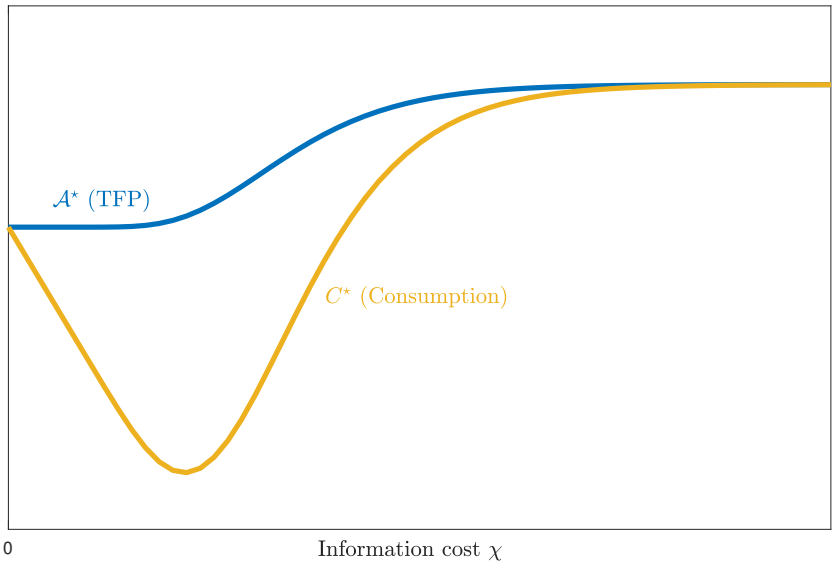
Proposition 5. *Let $\mu^\star$ be the marginal-type that is indifferent to producing information in equilibrium. Aggregate TFP and consumption in the rent-extracting economy then equal:*

$$\mathcal{A}^\star = \overline{\Psi}(\mu^\star) \cdot \mathcal{A}(\mu^\star, g^R) \quad \textit{and} \quad C^\star = \mathcal{A}^\star \cdot \left[N - \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right) \cdot \chi\right], \tag{47}$$

where the macro-level wedge satisfies:

$$\overline{\Psi}(\mu^\star) \equiv \left[\Psi(\underline{\tau}^\varsigma)^{\frac{\theta-1}{\theta}} \cdot (1-\zeta(\mu^\star)) + \Psi(\bar{\tau}^\varsigma)^{\frac{\theta-1}{\theta}} \cdot \zeta(\mu^\star)\right]^{\frac{\theta}{\theta-1}}, \tag{48}$$

Figure 7: Socially Harmful Advances in Data Processing



Note: The figure depicts the aggregate TFP and consumption, as they depend on the information cost χ, for the parametric case where segmentation is socially destructive (Definition 4).

where $\zeta(\mu^\star) \equiv \frac{g^R(\bar{\tau})^{\theta-1}\cdot\xi(\mu^\star)}{g^R(\underline{\tau})^{\theta-1}\cdot(1-\xi(\mu^\star))+g^R(\bar{\tau})^{\theta-1}\cdot\xi(\mu^\star)} \in (0,1)$ *is decreasing in* $\mu^\star$, $\mathcal{A}(\cdot)$ *and* $\xi(\cdot)$ *are defined in Lemma 1, and* $\Psi(\cdot)$ *and* $g^R(\cdot)$ *are defined by Equations (45) and (46), respectively.*

Contrasting Proposition 5 with its counterpart for the baseline economy (Proposition 2), we see that aggregate TFP and welfare in the rent-extracting economy are distorted by a macro-level wedge, $\overline{\Psi}(\mu^\star)$, which is the aggregate (i.e., firm-size weighted) counterpart to the firm-level wedge $\Psi_j \equiv \delta_j^S/\delta_j^R$ between profit and surplus maximization. This macro-level wedge reflects a pecuniary externality induced by firms' rent-extracting behavior. Each firm chooses its information, input, and pricing strategies to maximize private rents, taking aggregate outcomes as given. When such behavior is pervasive, however, it distorts equilibrium allocations and reduces overall efficiency and consumption. Since consumers are the ultimate owners of firms, this distortion thereby generates a wedge between what "private incentives" and "collective welfare" would dictate.[26] Crucially, this externality can be sufficiently strong that advances in data-processing technologies—which were unambiguously beneficial in the baseline economy (Proposition 3)—now become harmful in the aggregate:

Corollary 4. *An improvement in data-processing technologies, such as a decline in* χ *or an increase in the accuracy of information about consumer-specific tastes* $\bar{\tau}^\varsigma$*, raises the share of*

[26] A related pecuniary externality arises in monopolistic settings à la Dixit and Stiglitz (1977), where firms underproduce to extract rents from a representative consumer. A key distinction in our setting is that the magnitude of the externality depends endogenously on firms' information production.

information producers, $\Phi(-\mu^\star\sqrt{\tau_\mu})$*, but can reduce aggregate TFP,* $\mathcal{A}^\star$*, and consumption,* $C^\star$*.*

The intuition behind this counterintuitive result can be seen through a simple example, illustrated in Figure 7. Suppose that information production conveys little additional information about firms' productivity and demand states υ_j and ω_j, so that $\bar{\tau}^\upsilon \approx \underline{\tau}^\upsilon$ and $\bar{\tau}^\omega \approx \underline{\tau}^\omega$. Instead, information production concerns only consumer-specific tastes, $\{\varsigma_{ij}\}_i$. Consider a reduction in the cost of information, χ, from a prohibitively high level to zero, under parameter values for which segmentation is socially destructive, i.e., $\delta^S(\bar{\tau}^\varsigma) < \delta^S(\underline{\tau}^\varsigma)$ as in Definition 4. In this environment, the decline in consumer surplus induced by finer segmentation must more than offset the increase in firm profits, leading aggregate TFP and consumption to fall.[27]

Combined, these results show that improvements in data-processing technologies can, in some cases, amplify the distortions created by firms' rent-extracting behavior and thereby reduce overall welfare. Importantly, these welfare losses stem from misaligned incentives rather than from the data-processing advances themselves. With appropriate policy intervention, improvements in data processing would instead always raise aggregate efficiency and welfare.

5.5 Inefficiencies and Optimal Data Regulation

Before turning to the quantification of our baseline and extended frameworks, we briefly turn to the normative properties of the rent-extracting economy. To start with, we consider the problem of a social planner who maximizes aggregate welfare, $\int_i \mathcal{U}_i \cdot di$, by making optimal production and consumption choices, $\{x_j, n_j, \iota_j\}_j$ and $\{c_{ij}\}_{ij}$, acting under under the same technological and informational constraints as agents in the decentralized economy. We refer the reader to Appendix C.3 for the details of the planning problem.

We say that the decentralized equilibrium is *efficient* if its allocations coincide with those of the planner; otherwise, it is *inefficient*. We now state our first normative result:

Proposition 6. *The social planner's allocations coincide with those of the baseline economy. Hence, the laissez-faire equilibrium of the rent-extracting economy is generically inefficient.*

Proposition 6 establishes two key results. First, the equilibrium of our baseline economy is efficient. This result aligns with the well-established normative properties of CES economies in the tradition of Dixit and Stiglitz (1977) when factors of production are inelastic. Importantly, the additional features of our framework—information production and preference heterogeneity—do not alter this fundamental property.

Second, a direct implication of Proposition 6 is that the inefficiency of the rent-extracting economy stems from the misalignment between profit and social surplus maximization (Section

[27] A rise in $\bar{\tau}^\varsigma$ from a value close to $\underline{\tau}^\varsigma$ to a higher level yields qualitatively similar effects.

5.3). Indeed, if firms were to design trading mechanisms with the objective of maximizing social surplus rather than profits, the resulting allocations would be efficient.

Even though Proposition 6 serves as a useful normative benchmark, it also underscores a practical challenge in achieving an efficient outcome. Implementing this outcome may require *direct policy interventions* in firms' pricing strategies—an approach that may be challenging to implement in practice. Rather than pursuing such interventions, we focus in the following on more constrained (and arguably more realistic) policies. In particular, we explore interventions that target firms' information choices, which we broadly refer to as *data regulation policies*.

Specifically, we consider two sets of policy instruments. First, the planner can impose a tax T (in labor units) on information production, which affects equilibrium only through the determination of the marginal type $\mu^\star$.[28] Second, the planner can directly restrict firms' access to information by reducing the precision of the signals they observe through information production. Formally, the planner may *garble* signals about consumer-specific tastes by lowering their precision from $\bar{\tau}^\varsigma$ to $z^\varsigma \cdot \underline{\tau}^\varsigma + (1 - z^\varsigma) \cdot \bar{\tau}^\varsigma$, where $z^\varsigma \in [0, 1]$ is a policy parameter.[29]

Taken together, these instruments—one non-targeted and the other more targeted—provide a parsimonious representation of data-regulation frameworks such as the General Data Protection Regulation (GDPR) or the Digital Services Act (DSA) in the EU, which aim to: (i) raise the cost of collecting, storing, and exploiting consumer data, or (ii) restrict how such information can be used even when data collection itself remains feasible.[30]

Proposition 7. *The optimal data regulation has the following features:*

1. *If segmentation is socially destructive, then the policy sets $z^\varsigma = 1$ and $T = 0$.*
2. *If segmentation is socially valuable, then the policy sets $z^\varsigma = 0$ and: (i) $T > 0$ when segmentation is severe, but (ii) $T < 0$ when segmentation is mild.*

Consider first the case in which segmentation is socially destructive (Definition 4). In this case, more precise information about consumer tastes reduces the surplus created by a firm. Firms do not just reallocate surplus away from consumers; by sharpening segmentation, they also worsen allocative distortions and destroy surplus. Social efficiency then calls for directly

[28] The information cost faced by a firm becomes $\chi + T$, and tax revenues are rebated lump sum to consumers.

[29] Since firms would never garble their own signals, $\bar{\tau}^\varsigma$ can be interpreted as a technological upper bound on data-processing capabilities. Thus, while regulation cannot expand this frontier, it can limit firms' ability to exploit it. Moreover, as we show in the Appendix, garbling signals about the aggregate states υ_j and ω_j is not optimal for the planner, as these sources of information are benign.

[30] See https://gdpr.eu and https://digital-strategy.ec.europa.eu/en/policies/digital-services-act for details on the GDPR and the DSA respectively. In the model, all information can be interpreted as information about consumers: signals about aggregates υ_j and ω_j summarize features of the distribution of consumer preferences (see footnote 11), while signals about ς capture consumer-level heterogeneity. The relevant distinction is therefore between common and consumer-specific information.

limiting the accuracy of consumer information. Garbling such information eliminates the inefficiency—formally, the macro-level wedge $\overline{\Psi}(\mu^\star)$ becomes independent of $\mu^\star$—and leaves no role for an information tax. Accordingly, the planner sets $z^\varsigma = 1$ and $T = 0$.

Consider next the case in which segmentation is socially valuable. Here, all sources of information are benign, and it is therefore never optimal for the planner to garble firms' signals. Nonetheless, the equilibrium level of information production may still be inefficient. The direction of this inefficiency depends on how the macro-level wedge, $\overline{\Psi}(\mu^\star)$, varies with the level of information production, $\mu^\star$. For example, when segmentation is severe, the wedge declines as firms acquire more information. In this case, part of the resulting increase in firm profits reflects a redistribution from consumers to firms rather than a genuine efficiency gain. Consequently, firms' private return to information exceeds its social return, leading to excessive information production. The planner corrects this distortion by setting $T > 0$.

Policy implications. The data-regulation policies characterized above speak directly to current policy debates about firms' use of consumer data. A central concern is that unrestricted data collection may facilitate discriminatory pricing. In response, regulatory frameworks such as the EU's GDPR and DSA impose broad restrictions on the collection, processing, and use of consumer data. While highly stylized, the policies analyzed in Proposition 7 capture key features of these approaches and clarify the conditions under which they can improve welfare.

First, our analysis shows that data regulation need not always restrict information production. When segmentation is socially valuable and mild, information production raises overall surplus, but firms do not fully internalize its economy-wide benefits. As a result, the laissez-faire leads to underinvestment in information. This observation cautions against regulatory approaches that treat all forms of data-driven segmentation as inherently harmful.

Second, when segmentation is socially destructive, broad-based restrictions on information production are blunt instruments. In this region, inefficiencies arise from specific types of information—namely, information about consumer-specific tastes—that directly exacerbate allocative distortions through pricing. Policies that selectively limit access to such information dominate uniform increases in the cost of data collection.

Finally, policy discussions often emphasize the potential harm that data-driven pricing may impose on consumer surplus. Our analysis instead highlights an efficiency rationale for intervention: in our setting, firms do not fully internalize how information production affects the total social surplus they generate. In the next section, we consider two polar benchmark environments—the baseline and the rent-extracting economy—which span the full range of firms' ability to extract consumer surplus, and use them to derive bounds on the aggregate implications of recent advances in data-processing technologies for the U.S. economy.

6 Supporting Evidence and Quantification

We have demonstrated how advances in data-processing technologies affect the economy by enabling firms to optimize their (i) scale of operations, (ii) product choice, and (iii) pricing strategies. Our analysis shows that improvements in data processing can, in general, be either *beneficial* or *detrimental* in the aggregate. In this section, we quantify the implications of these forces for the U.S. economy over the past two decades. Has the rise in firm informativeness, documented in Section 2, contributed to higher productivity and welfare? If so, by how much and through which channels? To answer these questions, we calibrate the model using U.S. firm-level data and evaluate the effects of advances in data-processing technologies in the two polar environments studied above—the baseline and the rent-extracting economy.

6.1 Model Validation and Parametrization

Before using our framework as a quantitative laboratory, we provide a first-pass validation of the model. Specifically, we document that the cross-sectional predictions of our theory are both *qualitatively* and *quantitatively* consistent with salient features of U.S. firm-level data.

Recall that the baseline and the rent-extracting economies generate similar firm-level predictions. Across both environments, more accurate information enables firms to allocate inputs more efficiently from a private perspective. As already mentioned, however, the two economies differ in their efficiency properties. In the baseline economy, information use is socially efficient. In the rent-extracting economy, by contrast, firms use information to extract surplus from consumers, potentially distorting allocations and reducing social surplus in the process. In this sense, the two environments represent opposite ends of a spectrum that varies firms' ability to extract consumer surplus. We exploit this property below to construct sharp bounds on the macroeconomic consequences of recent advances in data-processing technologies.

Qualitative Validation. We begin by validating the model's cross-sectional predictions, summarized in Corollary 3 and discussed in Sections 4.1 and 5.4.

1) Information and Resource Allocation: First, a central role of information in our framework is to improve the allocation of inputs *within* firms. Indeed, (revenue-based) measured tfp is, all else equal, higher and less volatile for more informed firms. Firms with more precise demand-side information are better able to tailor their products to consumer preferences, thereby increasing and stabilizing their effective demand. Consistent with this prediction, Panel (a) of Figure D.1 shows a pronounced negative relationship between a firm's squared revenue error and its measured tfp, even after controlling for firm age and sector. In both the data and the model, more accurate firms are, all else equal, more productive.[31] Figure D.2 further shows

[31] We estimate (revenue-based) firm-level tfp following Ottonello and Winberry (2020), modifying the ap-

that more accurate firms also exhibit lower volatility of measured tfp.

Second, another key role of information in our framework is to reduce (ex-post) factor misallocation *across* firms. It does so by enabling firms to better align their input choices with realized productivity shocks. Panel (b) of Figure D.1 reports cross-sectional dispersion in marginal revenue products, separately for informed and uninformed firms. We classify a firm as informed if (i) its mean-squared one-year-ahead revenue error is below the median, and (ii) it has at least three observations (Appendix D.1). Across multiple measures, informed firms exhibit systematically lower dispersion in marginal revenue products. Although distortions unrelated to information—such as those documented by David and Venkateswaran (2019)—likely account for a substantial share of overall dispersion, the 30-40 percent difference shown in Panel (b) is robust to alternative definitions of informed firms (e.g., top quartile of the error distribution), alternative industry classifications (e.g., six-digit NAICS), and over time. Overall, the evidence thus suggests that improvements in informational accuracy are systematically associated with more efficient input allocation across firms.

2) Information and Firm-Size Distribution: Finally, information and firm size are closely linked in our framework. Firms with higher ex-ante productivity—and therefore larger expected scale—choose to produce information. At the same time, information itself leads firms to expand by improving their overall efficiency. The model therefore generates a tight, bidirectional relationship between a firm's information and size. Consistent with this, Figure 2 shows that larger firms exhibit more accurate forecasts, all else equal. Table A.14 in the Appendix uses the panel structure of our dataset to show that firms with more accurate forecasts at time t subsequently experience faster growth, consistent with the model's predictions.

In sum, the main cross-sectional predictions of the theory—common to both the baseline and the rent-extracting cases—are qualitatively consistent with observed firm-level patterns. We next assess the model's quantitative fit, with particular emphasis on the bidirectional relationship between accuracy and size. Doing so requires us to parameterize the model.

Calibration. Our parameterization is designed to ensure that the model replicates key features of firm-level outcomes while capturing the rich heterogeneity in expectations.

1) Baseline Economy: To this end—in the baseline economy—we set the elasticity of substitution between goods, θ, equal to 3, following Hsieh and Klenow (2009), normalize the labor endowment, N, to 1, and calibrate the remaining parameters internally.

We calibrate the variance of ex-ante log productivity, τ_μ^{-1}, the variance of productivity shocks, τ_a^{-1}, and the parameters $\bar{\sigma}$ and $\underline{\sigma}$ governing the common component of demand. We

proach to account for a finite elasticity of substitution. This estimation is consistent with an extension of our framework that incorporates capital in production (Section 6.3).

assume the latter is symmetric so that $\bar{\sigma} = \exp(\widehat{\sigma})$ and $\underline{\sigma} = \exp(-\widehat{\sigma})$ for some $\widehat{\sigma}$. This normalization is without loss of generality, as it only rescales the level of aggregate TFP. These parameters are chosen to match the unconditional variance of log productivity in the first five years of our sample (2002–2007), as well as the conditional variance of productivity for informed and uninformed firms over the same period. Consistent with evidence from Brynjolfsson and McElheran (2016) on the adoption of data-driven decision-making, we conservatively assume that 10 percent of firms are informed in 2002.[32] The information cost parameter, χ, is calibrated to match this share. The precision of firms' information, $\{\underline{\boldsymbol{\tau}}, \bar{\boldsymbol{\tau}}\}$, is set to match the mean-squared error of revenue expectations for informed and uninformed firms in the initial period. Finally, we normalize the parameters governing the idiosyncratic component of demand and set $\gamma = 0.5$ and $h = \ell = 1$, as these also only rescale TFP; as information about consumers' idiosyncratic tastes plays no role in the baseline economy, we set $\bar{\tau}^{\varsigma} = \underline{\tau}^{\varsigma} = 0.5$.

2) Rent-extracting Economy: We calibrate the rent-extracting economy analogously to the baseline, except for the parameters governing firms' pricing decisions: the idiosyncratic demand parameters (γ, h, ℓ) and the precision of information about idiosyncratic tastes, $(\bar{\tau}^{\varsigma}, \underline{\tau}^{\varsigma})$.

Consistent with the evidence in Adams *et al.* (2025), we assume that data-driven discriminatory pricing was limited in the early 2000s, implying $\bar{\tau}_{\varsigma} = \underline{\tau}_{\varsigma} = 0.5$ at the beginning of our sample. As with the common demand component, we assume without loss of generality that the idiosyncratic demand component is symmetric: $h = \exp(\widehat{\varsigma})$, and $\ell = \exp(-\widehat{\varsigma})$ for some $\widehat{\varsigma}$. We calibrate $\widehat{\varsigma}$ and the share of high-demand consumers, γ, to match the evidence on non-linear pricing in Bornstein and Peter (2024). Specifically, we use their estimates based on Nielsen Retail Scanner Data that relate the log of average payments to the log of quantities. We compare the model-implied relationship to estimates from higher value-added items, where data quality is better, and match the number of firm-year observations to that in our I/B/E/S–Compustat sample. While the aggregate economy spans sectors beyond retail and price variation may reflect factors other than heterogeneous preferences, these estimates suggest that distortions arising from non-linear pricing could be substantial.

Table II reports the model-data moment match for both the baseline and the rent-extracting economies, while Table D.1 summarizes the calibrated parameter values. We note that the moment match is identical for the two economies, as over the initial sample the rent-extracting economy is merely a rescaled version of the baseline economy.[33]

Quantitative Validation: A central implication of our theory is a tight relationship between firms' information and size. Above, we have documented *qualitative* evidence consistent with

[32] We define an informed firm as one in the bottom 10% of the squared-error distribution.

[33] The reason is that the revenue-based demand shifter δ_j^R in the rent-extracting economy coincides with the demand shifter δ_j in the baseline, whereas the wedge $\Psi_j = \delta_j^S / \delta_j^R$ is the same for all firms since $\bar{\tau}^{\varsigma} = \underline{\tau}^{\varsigma}$.

Table II: Parametrization: Model vs. Data (2002-2007)

	Data	Baseline	Rent-extracting
Mean of log-productivity of uninformed firms	-0.001	-0.001	-0.001
Mean of log-productivity of informed firms	0.014	0.018	0.018
Unconditional variance of log-productivity	0.068	0.067	0.067
Variance of log-productivity of uninformed firms	0.071	0.067	0.067
Variance of log-productivity of informed firms	0.050	0.060	0.060
Root-mean-squared error of uninformed firms	0.141	0.141	0.141
Root-mean-squared error of informed firms	0.032	0.032	0.032
Share of information-producing firms	0.100	0.100	0.100
Bornstein and Peter (2004) coefficient	-0.390	.	-0.338*

Note: The table compares data moments from I/B/E/S-Compustat sample over the period 2002-2007 to those from the calibrated model frameworks. The table shows the mean and variance of log-productivity, in addition to the root-mean-squared-error of firms' one-year-ahead log-revenue forecasts. Firm productivity is estimated as in Ottonello and Winberry (2020). We define an informed firm as a firm that is in the bottom 10 percent of the mean-squared-error distribution over the initial period and for which we have at least 3 observations. The Bornstein and Peter (2024) regression coefficient of the relationship between log average payments for an item and the corresponding quantity only pertain to the rent-extracting economy.

this prediction. We now use our calibrated model to also *quantitatively* validate our framework.

To do so, we simulate expectation errors over a 20-year period, linearly reducing the information cost parameter, χ, to replicate the 41% increase in average accuracy documented in Section 2. At the same time, we increase $\bar{\boldsymbol{\tau}}$ to capture improvements in data processing and the associated decline in tfp volatility among information producers. In the rent-extracting case, where the precision of information about consumer-specific tastes also affects firm-level tfp, we increase $\bar{\tau}_{\varsigma}$ to match the differential growth in tfp between informed and uninformed firms. The number of firm-year observations in each case matches that in the data.

Table III and Figure D.3 show that both calibrated economies generate a negative relationship between firm errors and firm size.[34] Even though our calibration does not explicitly target the size-accuracy relationship, both economies generate a pattern that closely mirrors the data. The regression coefficient on size—measured by a firm's employment quintile—is -0.39 in the baseline economy, -0.41 in the rent-extracting economy, and -0.42 in the data (Table III). In both models and in the data, moving up one employment quintile reduces the

[34]To match the patterns in the data, we estimate that:

- (Baseline economy) The cost of information production, χ, declined by 37%, while the precisions of produced information, $\bar{\tau}^{\upsilon}$, $\bar{\tau}^{\omega}$, and $\bar{\tau}^{\varsigma}$, increased by 10%, 1%, and 0%, respectively.
- (Rent-extracting economy) The cost of information production, χ, declined by 27%, while the precisions of produced information, $\bar{\tau}^{\upsilon}$, $\bar{\tau}^{\omega}$, and $\bar{\tau}^{\varsigma}$, increased by 10%, 1%, and 30%, respectively.

Table III: Size and Accuracy Relationship

	log. squared errors		
	Data	Baseline	Rent-extracting
Size (labor)	-0.42^{***}	-0.39^{***}	-0.41^{***}
	(0.05)	(0.01)	(0.01)
Time (years)	0.01	-0.01^{***}	-0.02^{***}
	(0.01)	(0.01)	(0.01)

Note: Least-squares estimates of the relationship between squared normalized errors and firm size (quintiles of the initial employment distribution). We estimate this relationship both in the I/B/E/S-Compustat data and in the calibrated model. The column labeled data further controls for firm fixed effects, log revenue vol., and age (Column (4) in Table I). Robust (clustered) standard errors in parentheses. *p<0.1,**p<0.05,***p<0.01.

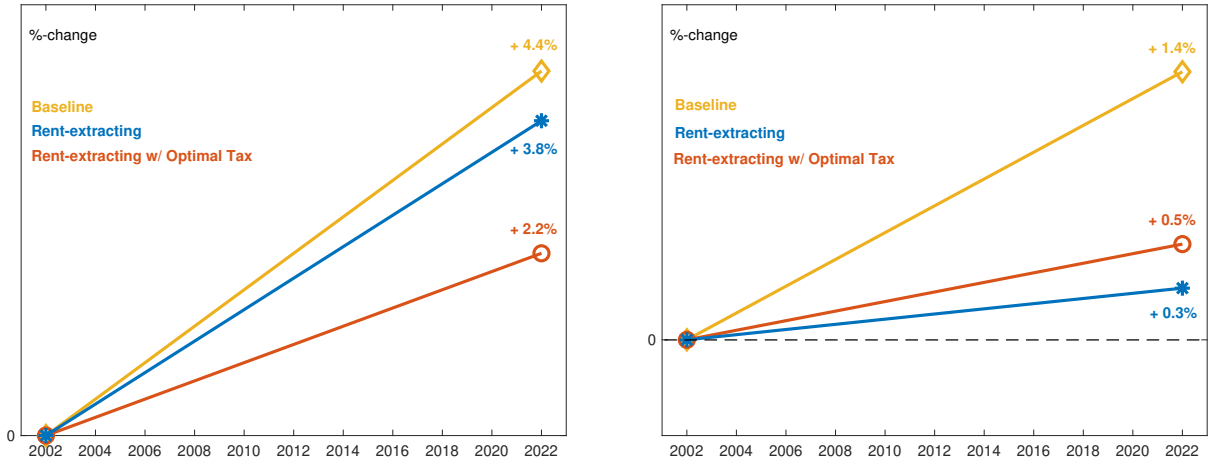
squared error by roughly 40 percent of its mean value. Both models further replicate the slope of the size-accuracy relationship reasonably well (Figure D.3), although they somewhat understate the improvement between the first and second quintiles.

Crucially, the two calibrated environments also replicate key patterns in the evolution of data use and data expenditures over time. First, while precise estimates of firms' adoption of data-processing technologies remain difficult to obtain (Baley and Veldkamp, 2025), recent work by Brynjolfsson and McElheran (2024) estimates that approximately 73% of medium-to-large manufacturing firms systematically used data to inform their economic decision-making by 2021. Our baseline and rent-extracting frameworks slightly overstate this figure, predicting that 80% and 77% of firms, respectively, were "information producers" by the end of our sample (2022). Nevertheless, both the magnitude and the trajectory of the increase are remarkably similar—particularly given the framework's minimalist nature.

Second, the calibrated environments also match firms' overall expenditures on data reasonably well. The calibrated models imply that the average firm spends around 1.9–2.0% of its revenue on data by the end of the sample—somewhat above the 1.5% reported by the IDC IT Wallet Survey, which covers hardware, software, and cloud-computing expenditures for publicly-listed U.S. firms in 2021 (IDC, 2021). This modest gap is however expected, as the framework is designed to capture a broader set of data-related expenditures.

Finally, the relationship between firms' data expenditures and their accuracy in the two environments is also consistent with that observed in the I/B/E/S-Compustat-IDC merger. Table A.13 shows that the empirical relationship between data spending and accuracy closely matches the model-implied relationship between expenditures on information production and accuracy generated by the calibrated baseline and rent-extracting environments. This correspondence is reassuring, as it suggests that our model not only matches observed data

Figure 8: Effects of Improvements in Data-Processing Technologies



(a) TFP (b) Welfare

Note: Panel (a) shows the effects for the calibrated scenarios on TFP, $\mathcal{A}$. Panel (b) depicts the results for overall welfare, C. We showcase results for the baseline and the rent-extracting economies. We also illustrate the results for the rent-extracting economy under the optimal data regulation described in Proposition 7.

expenditures but also captures the firm-level returns to information.

In summary, the results in this subsection show that the model both *qualitatively* and *quantitatively* captures salient relationships between the accuracy of firms' expectations and firm-level outcomes. We therefore conclude that the model provides a suitable laboratory to study the macroeconomic consequences of advances in data processing.

6.2 Quantification and Decomposition

We quantify the economy-wide effects of the estimated declines in information costs, χ, and increases in information precision, $\bar{\boldsymbol{\tau}}$. Figure 8 and Table IV summarize our results.

On balance, the estimated effects of advances in data-processing technologies are substantial. Over the past two decades, economy-wide TFP gains attributable to advances in data processing are estimated to be 3.8-4.4% (Panel (a) of Figure 8), depending on the environment considered (i.e. rent-extracting vs. baseline). For comparison, TFP in the data has increased by approximately 15% over our sample period.[35] Importantly, our estimates furthermore account for roughly one-half to three-quarters of the potential TFP gains that could be realized from eliminating information frictions.[36] Although there are clear uncertainties surrounding

[35] See, e.g., the FRED series on TFP (`https://fred.stlouisfed.org/series/RTFPNAUSA632NRUG`).

[36] We compute this benchmark by setting the cost of information production to zero and increasing information precision toward infinity and one for productivity- and demand-side information, respectively. We then recompute equilibrium outcomes in the economy without information frictions. The resulting increase in aggregate TFP is 8.8% and 5.4% in the two environments, respectively.

Table IV: Decomposition of The Rise in TFP

Model	Overall (%)	Scale (pp)	Product (pp)	Pricing (pp)	$\underline{\Psi}$	$\bar{\Psi}$
Baseline	4.4	2.1	2.3	0	.	.
Rent-extracting	3.8	2.0	3.1	-1.3	1.07	1.05

Note: The table decomposes the TFP rise in Figure 8 into its three constituent channels: (i) scale, (ii) product design, and (iii) pricing. The table does so for both the baseline and the rent-extracting economies.

these first-pass estimates, taken at face value, our results point to two findings. First, improvements in firms' use of data have been a meaningful contributor to the rise in TFP over the past two decades. Second, perhaps more surprisingly, a large share of the potential productivity gains from advances in data processing may have already been realized.

Table IV decomposes the overall rise in TFP into its three constituent channels, highlighting a key strength of our framework.[37] The results show that firms' improved ability to determine their optimal scale of operations and product choice contribute roughly equally to TFP growth over the sample period: 2.1 pp and 2.3 pp, respectively, in the baseline economy, and 2.0 pp and 3.1 pp, respectively, in the rent-extracting economy. However, crucially, notice that in the rent-extracting economy the gains from either of these channels is offset by firms' distorted incentives to produce information to extract rents from consumers. We estimate the latter, resulting drag on TFP to be as much as –1.3 pp over the sample period.

Although the overall rise in TFP is similar across the baseline and the rent-extracting economies (4.4% vs. 3.8%), the estimated welfare consequences differ sharply. Panel (b) of Figure 8 shows that in the baseline economy overall welfare increases by approximately 1.4% as a result of advances in data processing over the sample period. In contrast, under the rent-extracting scenario, these potential welfare gains are almost entirely offset by firms' excessive information production, resulting in an estimated welfare increase of only 0.3%.

Even though the rent-extracting economy may overstate firms' ability to extract consumer surplus in a distortionary manner (see, however, Section 6.3), the results nevertheless demonstrate that a substantial disconnect between TFP gains from advances in data-processing technologies and their welfare consequences is quantitatively possible. Once one accounts for endogenous changes in firms' pricing behavior, the welfare benefits of these advances may be significantly attenuated—even in the presence of sizable productivity gains.

This disconnect raises the question of the potential welfare improvements that corrective policy can deliver. Table IV implies that the rent-extracting economy operates in the pa-

[37] This decomposition follows directly from Proposition 5.

rameter region in which segmentation is *socially valuable* but *severe* (Definition 4). As we have shown in Section 5.5, this creates a clear rationale for corrective data regulation that discourages excessive information production by firms (Proposition 7). Panel (b) of Figure 8 shows that the optimal information tax—that aligns firms' incentives with collective welfare—increases the welfare gains due to advances in data processing from 0.3% to 0.5%. Achieving these gains requires a substantial reduction in information production: the share of information-producing firms at the end of the sample falls by roughly 30 pp, to around 50%.

Finally, a strength of our framework is that it can speak to the contribution of data advances to the rise in firm-size concentration over the past two decades. At the start of our sample, we estimate that information production increases the size of a firm (measured by employment n_i) by 10%, on average. At the end of our sample, this contribution has, all else equal, increased to 11% in the baseline and to 13% in the rent-extracting economy.[38] Because of the rapid rise in share of information producers, we estimate that advances in data-processing have therefore, all else equal, increased firm-size concentration (as measured by the 90/10 ratio) by 2-3%. In the I/B/E/S-Compustat sample, by contrast, this ratio has increased by 8.5%. Through the lens of our framework, advances in data-processing technologies have been an important driver of the rise in firm-size concentration.

6.3 Quantitative Refinements

Capital and Variety Accumulation. A potential concern with the above estimates is that they may be *conservative*. In the framework studied so far, both the supply of production factors and the set of available varieties are fixed, which limits the economy's adjustment to improvements in data processing. To assess the implications in a more elastic environment, we extend both the baseline and the rent-extracting economies to allow for capital accumulation and endogenous variety creation. For tractability, we study this extension in steady state, and we relegate its details to Appendix D.6. Figure 9 compares the effects of improvements in data-processing technologies in this extended model with those reported in Figure 8, using the same calibration strategy.[39] In this extension, information-production costs are measured

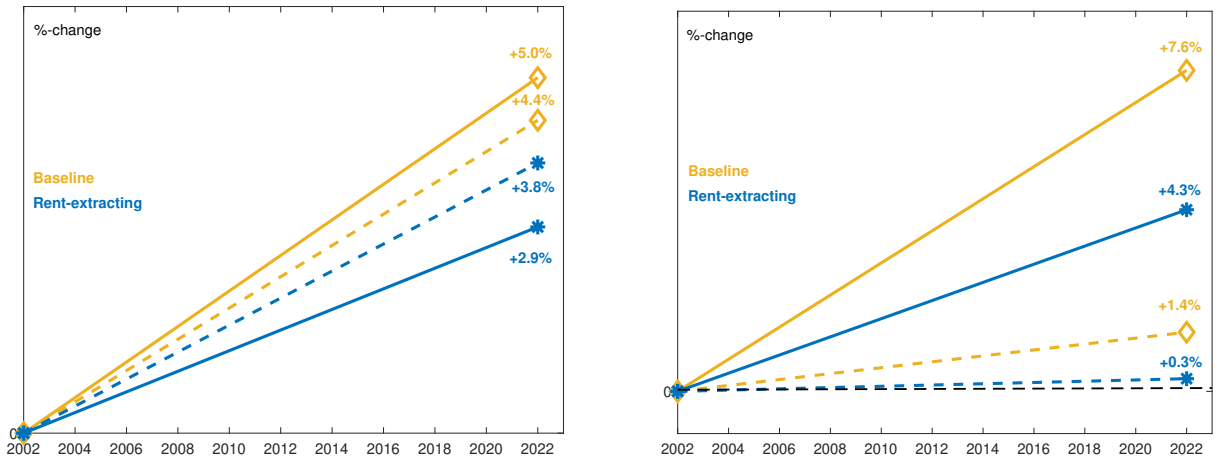
[38] Formally, in the baseline economy, the ratio of expected employment of an informed to uninformed firm can be shown to equal:

$$\Delta S = \frac{\frac{\int_{\mu_j > \mu^\star} e^{(\theta-1)\cdot\mu_j}\cdot dj}{\mathbb{P}(\mu_j > \mu^\star)}}{\frac{\int_{\mu_j < \mu^\star} e^{(\theta-1)\cdot\mu_j}\cdot dj}{\mathbb{P}(\mu_j < \mu^\star)}} \cdot \left(\frac{g(\bar{\boldsymbol{\tau}})}{g(\underline{\boldsymbol{\tau}})}\right)^{\theta-1}, \tag{49}$$

which captures both firm selection into information production and effect of information on firm size. In the rent-extracting economy, the information shifter $g(\cdot)$ is simply replaced by the modified information shifter $g^R(\cdot)$. This implies that the contribution of information to the percent rise in firm size is $1 - (g(\underline{\boldsymbol{\tau}})/g(\bar{\boldsymbol{\tau}}))^{\theta-1}$.

[39] In the extended model, we estimate that the cost of information production, χ, falls by 32% in the baseline economy and by 21% in the rent-extracting economy. The upper bounds on the precision parameters

Figure 9: Capital and Variety Accumulation



(a) TFP (b) Welfare

Note: Panel (a) shows the estimated effects for the baseline and the rent-extracting economies on (steady-state) TFP, $\mathcal{A}$. Solid lines depict results from the augmented quantification, while dashed lines depict those from Figure 8. Panel (b) shows the results for (steady-state) consumption, C.

in units of output and are therefore naturally interpreted as a form of intangible investment.

Relative to our earlier estimates, the gap between the baseline and rent-extracting scenarios widens somewhat in the extended model. While the estimated TFP increase in the baseline scenario remains similar (5.0% compared to 4.4%), the gain in the rent-extracting case is smaller (2.9% versus 3.8% previously). As shown in Table D.2 in the Appendix, this increased difference is mainly driven by the TFP drag associated with firms' discriminatory pricing behavior. In equilibrium, the distortions that arise from firms' pricing behavior reduce the economy's effective market size, leading to a gradual decline in the mass of varieties supplied. Over time, this general-equilibrium response further depresses aggregate TFP, as the contraction in product varieties amplifies the adverse consequences of rent extraction.

Finally, although the notion of welfare is more complex in this dynamic extension—particularly given that the transition to a new steady state may take time—Figure 9 shows that the estimated increase in steady-state consumption is now larger. This reflects the economy's increased ability to accumulate production factors (i.e., capital), which amplifies the consumption response to improvements in data-processing technologies. Importantly, however, a sizable gap between the baseline and rent-extracting scenarios persists: the steady-state consumption gain in the rent-extracting case is less than 2/3rds of that in the baseline. This

$\bar{\tau}^{\upsilon}$ and $\bar{\tau}^{\omega}$ increase by 10% and 1%, respectively, in both cases, while $\bar{\tau}^{\varsigma}$ rises by 27% in the rent-extracting economy. Appendix D.6.2 discusses the calibration of the extended model in detail. Relative to the static model, the estimated decline in χ is smaller, reflecting the greater responsiveness of the extended economy to improvements in information.

Table V: Alternative Calibration

Parameters	TFP (%)	Welfare (%)
All parameters (A)	(4.3, 3.6)	(1.5, 0.5)
Non-information (B)	(0.7, 0.5)	(0.5, 0.4)
Information ($C = A - B$)	(3.5, 3.1)	(1.0, 0.1)

Note: The table shows the effects of (A) changing both information and productivity parameters, as described in the text (i.e., "all parameters"); (B) changing only productivity parameters (i.e., "non-information parameters": $(\tau_\mu, \tau_a, \widehat{\sigma}, \widehat{\varsigma})$); and $(C = A - B)$ the difference between the two cases, attributable to changes in information parameters. The table shows the results both for the baseline (first) and the rent-extracting (second) economies identified using the calibration that changes both information and productivity parameters.

underscores once again the potential welfare gains from corrective data regulation.

Re-calibrated Productivity Parameters. Our baseline estimates assess the impact of advances in data-processing while holding fixed the underlying process for firm-level productivity (Equation (4)). However, due to the two-sided relationship between firm size and information in our framework, a potential concern is that the productivity process itself may have evolved over time, thereby inducing changes in information production and the accuracy of expectations even absent advances in data processing. To explore this possibility, Table V reports results from an alternative calibration that—alongside recalibrating the information parameters—also recalibrates the parameters governing the productivity process, (τ_μ, τ_a), as well as those governing consumer demand, $(\widehat{\sigma}, \widehat{\varsigma})$.[40] For both the baseline and the rent-extracting scenarios, the contribution of data-processing technologies to TFP remains similar to before: TFP gains are once more estimated to be in the range of 3-4%. Similarly, the discrepancy between the welfare gains in the baseline and the rent-extracting scenarios is likewise still substantial. Indeed, welfare is estimated to have hardly risen in the recalibrated rent-extracting economy (0.1%). This underscores the robustness of our findings.

To summarize, calibrating the model to firm-level developments over the past two decades, we estimate that advances in data processing have led to a substantial increase in TFP (c. 3.0-5.0%). Yet our analysis also cautions that the associated welfare gains may have been more limited. In the baseline economy, where information production is efficient, welfare gains are meaningful (c. 1.0-7.5%). By contrast, in the rent-extracting economy they are considerably smaller (c. 0.1-4.0%), due to excessive information production by firms that amplifies distortions arising from data-driven price discrimination. Our findings thus highlight a potential role for data regulation in mitigating losses from such pricing-related distortions.

[40] Specifically, we recalibrate $(\tau_\mu, \tau_a, \widehat{\sigma}, \widehat{\varsigma})$ at the end of the sample period to match the same firm-level moments targeted at the beginning of the sample. We use averages over the final three years as targets.

7 Conclusion

Advances in data-processing technologies hold the potential to transform many dimensions of economic life. In this paper, we have focused on one such dimension: the capacity of these technologies to enhance firms' information about economic fundamentals. Using micro data on managerial expectations, we have documented a systematic rise in the accuracy of U.S. firms' expectations over the past two decades, and showed that this increase is closely linked to shifts in the firm-size distribution and firms' data expenditures. To assess the macroeconomic implications of this trend, we developed a unifying quantitative framework in which information enables firms to optimize their *scale*, *product choice*, and *pricing strategies*.

Consistent with micro-level data, our model predicts that firms that use information more effectively allocate inputs more efficiently, design better products, and grow faster and larger over time. We have quantified the aggregate consequences of these micro-level shifts. For plausible parameters, our benchmark estimates suggest that TFP and household welfare would have been between 3.8–4.4% and 0.3–1.4% lower in 2022, respectively, absent the increase in the information accuracy of U.S. firms. We have further decomposed these overall effects and found that the TFP gains stem roughly equally from firms' ability to better determine their optimal scale of operation and their optimal product choice.

That said, our analysis also cautions that a substantial share of the welfare gains from higher TFP may have been offset by excessive information production. Our estimates suggest that firms may have devoted resources to information that—by facilitating greater rent extraction from consumers—have not translated into commensurate efficiency gains. The welfare effects of data-driven advances depend critically on how, and for what purposes, information is used. The design of an appropriate data regulation framework is thus central to ensuring that data-driven advances translate into broad-based welfare gains.

Finally, our framework opens several avenues for future research. A particularly promising direction, in our view, is to incorporate household-side responses—i.e., how consumers use data to navigate the product space and how this behavior interacts with firms' strategic choices. Another valuable extension would be to leverage product-level transaction data across a broader set of sectors to refine empirical estimates of the consumer-side costs associated with data-driven firm behavior. While this paper offers an initial step using firm-level data, much remains to be understood regarding the broader welfare implications of the data revolution.

A Motivating Evidence

A.1 Data Construction: I/B/E/S-Compustat

In our main analysis of firms' expectations, we use a combination of the I/B/E/S managerial guidance database and Compustat Fundamentals Annual. The combined sample for the I/B/E/S-Compustat merger covers the period 2002-2022 for 12,917 firm-years spanning 2,570 US firms. To construct our sample, we follow convention and discard utilities and financials, as well as any firm-years that have negative or non-existing values for revenue, employment, and/or the capital stock. We focus on revenue expectations, which comprise the lion's share of all forecasts provided by managers, and analyze one-year ahead annual expectations.[41] We only use "centered forecasts": that is, either point estimates or forecasts that are stated as a range. In the latter case, we use the mid-point of the range as the point estimate. We remove observations that are related to the top and bottom 1 percent of the error distribution.

Variable Definitions: We use the following variables from Compustat Fundamentals Annual: revenue (code: `sale`), profits (code: `ib`), capital (code: `ppent, ppegt`), investment (code: `capx`), assets (code: `at`), employment (code: `emp`), mergers and acquisitions (code: aqa), and industry classification (code: `naics` and `sic`). We measure a firm's debt as the total net value of liabilities (code: `dllt+dlc-che`), and the stock of acquired intangibles, adjusted for amortization and financial goodwill as in Chiavari and Goraya (2023) (code: `ITAN+AM-GDWL`). We deflate nominal variables where appropriate with US CPI (code: `CPIAUCLS` from FRED) and compute (revenue-based) total factor productivity as in Ottonello and Winberry (2020), adjusting for a finite constant degree of elasticity of substitution between goods. Finally, Compustat only has limited data on wages. We use total labor and related expenses as our measure of the overall wage bill (code: `xlr`). We link the Compustat data with the I/B/E/S database using the CRSP ID that is available for both. The annual expectations employed from the I/B/E/S managerial guidance database (code: `val1` and `val2`) are those that pertain to "centered forecasts" (code: `fdesc=1,2`) in millions or billions of USD. We mainly study expectations of future revenue (code: `measure=SAL`), although we also consider expectations of future profits (code: `measure=NET`) and capital expenditures (code: `measure=`cpx). We define a firm's error as the difference between the realized value of the variable from Compustat and the one-year-ahead expectation of the variable from I/B/E/S.

Descriptive Statistics: Table A.1 reports descriptive statistics for our merged data set.

[41] For an individual firm, we study the first forecast made in the year (Jan-April) that pertains to the firm's end-of-year financial results. Firms mainly report previous year's financial results in Q1 of the following year.

Table A.1: Descriptive statistics: I/B/E/S-Compustat

Variable Name	Obs.	Mean	Std.	Median
Revenue	12,917	3,762	11,518	769.49
Profits	12,917	279.30	1,403	28.98
Capital	12,825	1,052	4,692	122.79
Investment	12,910	184.13	857.47	46.70
Wages	845	1,841	3,974	353.89
Assets	12,917	5,720	22,152	1012.85
Employment	12,835	12.90	32.42	2.90
Revenue/capital	12,821	13.23	32.16	6.97
Expectation/capital	12,821	14.71	51.24	7.00
Expectation Log	12,821	1.94	1.13	1.95
Expectation Error	12,567	-0.14	1.91	0.01
Expectation Error Log	12,563	-0.01	0.13	0.00

Notes: The table reports descriptive statistics for the sample of 2,570 firms from 2002-2022 in the merged Compustat-I/B/E/S database. The units of the first seven rows are USD millions. The employment row is in '000-employees. The first eight rows capture, respectively, firm revenue, GAAP net-profits, book value of the capital stock, total value of capital expenditures, end-of-period total liabilities and assets, overall expenditures to labor and related expenses, and the total number of employees. The next three rows measure revenue scaled by a firm's tangible capital and the (log) of the year-ahead expectation. The final two rows are for the year-ahead error defined as realized future (log)-revenue minus (log of) the expectation. In the final two rows, observations have been removed that are in top and bottom 1 percent of the error distribution.

A.2 Data Construction: Duke-Richmond Fed CFO Survey

The CFO Survey is a quarterly survey of U.S. business leaders designed to elicit the financial outlook for their firms, the economic challenges they face, and their expectations about the broader U.S. economy. We exploit a combination of survey answers from The CFO Survey and data on economy-wide outcomes from FRED. The sample covers the period 2020-2022, the period for which data is available, for 3,470 firm-years spanning 826 U.S. firms. We remove expectations that are not one-year-ahead, as well as any firm-years that have non-existing profits. We throughout focus on annualized real GDP growth expectations (code: `GDPC1`). Firm size is measured by the number of domestic full-time employees. The size buckets used below correspond to quintiles of the 2020 size distribution: `size=1` (fewer than 6 employees); `size=2` (6-40 employees); `size=3` (40-130 employees); `size=4` (130-500 employees); and `size=5` ($>$ 500 employees). The familiarity with the concept of "Gross Domestic Product" (GDP) is measured numerically with a scale from 1-3.

A.3 Data Construction: IDC IT Wallet Database

The IDC IT Wallet Database is an annual database comprised of firms across the world and their expenditures on different data categories. To estimate these expenditures, IDC uses a proprietary combination of bottom-up (e.g., surveys) and top-down approaches (e.g., sales information from market leaders). We exploit estimates for publicly-traded U.S. firms expenditures on `hardware` (on-premise, off-premise, used for infrastructure and in systems), `software` (on-premise, off-premise, and in the cloud), and `IaaS` (cloud computing). `Overall data expenditures` are measured by the sum of `hardware`, `software`, and `IaaS`. Our sample covers the year 2020 for 6,466 firms. We merge this data set on firms' data expenditures with the I/B/E/S-Compustat sample described above in Appendix A.1.

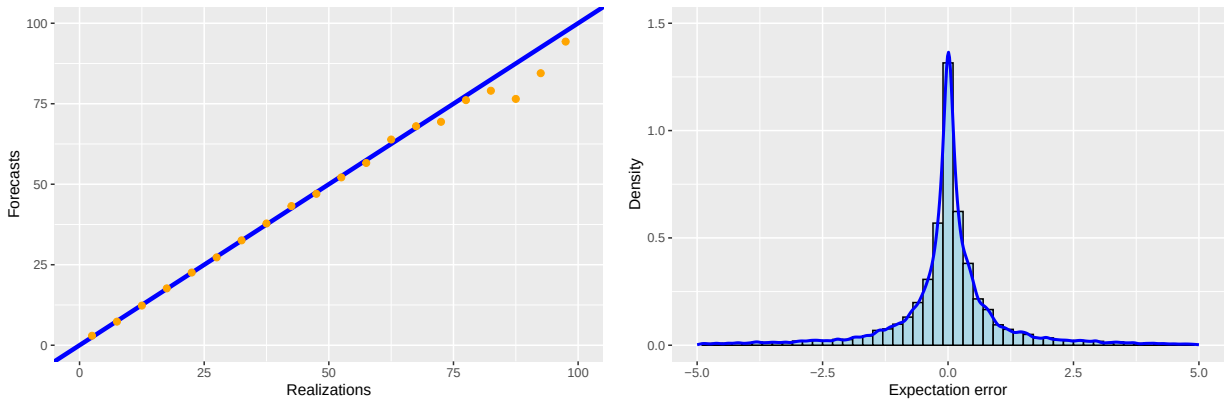
A.4 Additional Data Comments

In this appendix, we provide a brief overview of the relationship between firms' revenue expectations and their input choices in the I/B/E/S-Compustat database. First, notice that firms' revenue errors are close to unbiased. Table A.1 shows that the mean error of firms' log revenue errors is, for example, -0.01 compared to an average value of log revenue of 1.94. Figure A.1 provides a full bin-scatter plot of firms' revenue expectations and realizations, and shows that the associated error distribution is also close to symmetric. These results are consistent with the evidence in, e.g., Chen *et al.* (2023), who show that Japanese firms' expectations about their own sales, in addition to their expectations about macroeconomic and sector-specific inflation rates, are close to unbiased. All else equal, firms in our sample do not appear to systematically skew their revenue expectations one way or another. Relatedly, Chen *et al.* (2024) explore how positive and negative E/P/S revisions respond to new information in the I/B/E/S-Compustat database, finding also a consistent pattern. Second, Table A.2 conducts an exercise akin to that in Tanaka *et al.* (2020). Panel (a) documents the relationship between the realized growth in inputs and the current-period (firm-specific) expectation of *future revenue*. We find that, all else equal, more optimistic firms employ and invest more, consistent with these firms being viewed as more optimistic. The estimated effect sizes are, furthermore, substantial: a 1 percent increase in expected revenue is associated with 0.14 percent increase in investment, for example. In Section 6.1 in the main text, we discuss how the *accuracy* of firms' expectations, in addition to their level, systematically affect firms' input choices, consistent with our model framework. Panel (b) in Table A.2 instead explores the relationship between the realized growth in various inputs and the *previous period's revenue error*. We find that firms that have been positively surprised—i.e., have higher revenue than previously expected, and hence a positive revenue error—subsequently employ and invest more, in line with these firms being genuinely surprised about the revenue realization and also more opti-

mistic. We conclude that the results in Table A.2 are consistent with those in Tanaka *et al.* (2020), among others, who show that firms who are more optimistic about the future invest and employ more, and that positive profit (or revenue) surprises result in more inputs being employed and allocated subsequently.

Finally, we note that the arguments we provide in the main text do not require that firms' reported expectations strictly equal their (correct) mathematical expectation of future revenue. We do not require the complete absence of strategic or behavioral drivers of expectations. We only require that changes in reported expectations (and in their accuracy) in part reflect changes in information. The results in Tables A.1 and A.2 are consistent with this role of information. The results in Table A.5, which show that larger firms in the Duke-Richmond CFO survey report more accurate expectations of a variable (real GDP growth) over which they have no control, further bolster this case.

Figure A.1: Expectations, realizations, and errors



Panel (a): expectations and realizations

Panel (b): distribution of errors

Note: Data from I/B/E/S-Compustat. Panel (a): a bin-scatter plot of firms' one-year-ahead revenue expectations and their realizations. Panel (b): a histogram of the associated error distribution. Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Sample: 2002-2022.

Table A.2: Expectations and input choices

	Panel (a): outcomes and expectations		
	Employment (%)	Capital (%)	Investment (%)
Revenue expectation (%)	0.069***	0.222***	0.139*
	(0.026)	(0.085)	(0.078)
Firm age (quintile)	-2.072	-5.794	-0.014
	(1.627)	(4.198)	(1.725)
Observations	10,260	10,277	10,255
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
F statistic	52.856***	137.772***	26.116***
	Panel (b): outcomes and errors		
	Employment (%)	Capital (%)	Investment (%)
Revenue error lagged (%)	0.188***	0.195	0.661**
	(0.053)	(0.266)	(0.312)
Firm age (quintile)	-2.492	-7.159	-0.888
	(1.772)	(4.733)	(2.079)
Observations	10,020	10,096	10,069
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
F statistic	4.210**	3.228**	5.3786***

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the relationship between realized growth in employment (capital and investment) and the current-period firm-specific expectations of revenue growth. Panel (b): estimate of the relationship between realized growth in employment (capital and investment) and the one-period lagged revenue error. The table also controls for a firm's age, measured in quintiles of the overall age distribution. Expectation (errors) related to the top and bottom 1 percent of the error distribution have been removed. All estimates controls for time and firm fixed effects. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

A.5 Additional Estimates

Table A.3: Time evolution of accuracy and size

	Panel (a): revenue errors and time			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.024^{***}	-0.015^{***}	-0.028^{***}	-0.024^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	1.251^{***}	1.135^{***}	1.303^{***}	1.219^{***}
	(0.037)	(0.026)	(0.085)	(0.065)
Observations	12,567	12,563	12,567	12,563
Covid dummy	×	✓	×	✓
Residual std. error	1.835	1.278	4.302	3.148
F statistic	67.11^{***}	57.23^{***}	17.70^{***}	22.59^{***}
	Panel (b): size and time			
	50th perc.	70th perc.	80th perc.	90th perc.
	(1)	(2)	(3)	(4)
Time	0.011^{***}	0.012^{***}	0.009^{***}	0.005^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.567^{***}	0.374^{***}	0.215^{***}	0.111^{***}
	(0.015)	(0.016)	(0.007)	(0.005)
Observations	21	21	21	21
Residual std. error	0.029	0.032	0.021	0.014
F statistic	119.85^{***}	115.36^{***}	153.64^{***}	98.96^{***}

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead revenue errors on time. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b): estimate of the coefficient of the share of firms with employment greater than the xth percentile of firms in 2002 on time. In panel (a), columns (1) and (3) are in levels, whereas columns (2) and (4) pertain to the logs of variables. Robust standard errors in parentheses. Sample: 2002–2022.

Table A.4: Time evolution of accuracy: sector fixed effects

	Panel (a): sector fixed effects			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.024^{***}	-0.016^{***}	-0.027^{***}	-0.026^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	0.586^{***}	1.818^{**}	0.405^{***}	3.472
	(0.117)	(0.860)	(0.106)	(2.738)
Observations	12,567	12,563	12,567	12,563
Covid dummy	✓	✓	✓	✓
Sector FE	✓	✓	✓	✓
Residual std. error	1.814	1.269	4.282	3.141
F statistic	17.32^{***}	13.71^{***}	7.15^{***}	5.66^{***}
	Panel (b): sector×time fixed effects			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.023^{***}	-0.011^{***}	-0.028^{***}	-0.020^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	1.246^{***}	1.114^{***}	1.301^{***}	1.209^{***}
	(0.033)	(0.026)	(0.082)	(0.067)
Observations	12,567	12,563	12,567	12,563
Sector×time FE	✓	✓	✓	✓
Residual std. error	1.652	1.257	4.184	3.176
F statistic	79.09^{***}	29.37^{***}	18.52^{***}	15.54^{***}

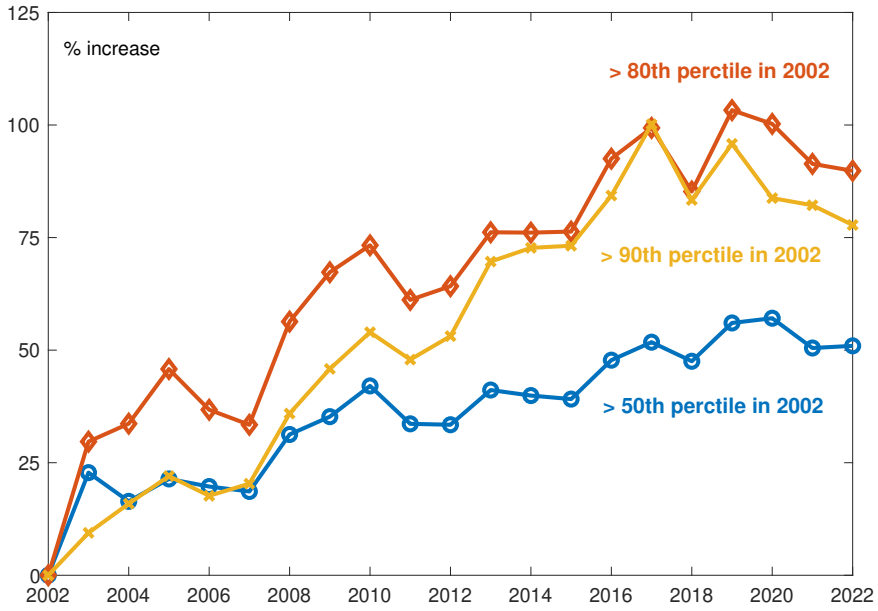
Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead revenue errors on time after having partialled out for sector (NAICS-2) fixed effects and a COVID dummy. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b) instead partials out for sector×time fixed effects. Columns (1) and (3) are in levels, whereas Columns (2) and (4) pertain to the logs of variables. Robust standard errors in parentheses. Sample: 2002–2022.

Table A.5: Output expectations from the Duke CFO Survey

	Squared error		Absolute error	
	(1)	(2)	(3)	(4)
Firm size	-0.077^{**}	-0.060^{*}	-0.050^{***}	-0.045^{***}
	(0.028)	(0.024)	(0.012)	(0.010)
GDP familiarity		0.033		0.018
		(0.043)		(0.025)
Constant	1.639^{***}	1.676^{***}	1.648^{***}	1.681^{***}
	(0.115)	(0.090)	(0.051)	(0.042)
Observations	1,584	1,464	1,584	1,464
Sector FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Residual std. error	1.578	1.544	0.812	0.793
F Statistic	19.774^{***}	20.165^{***}	29.175^{***}	29.814^{***}

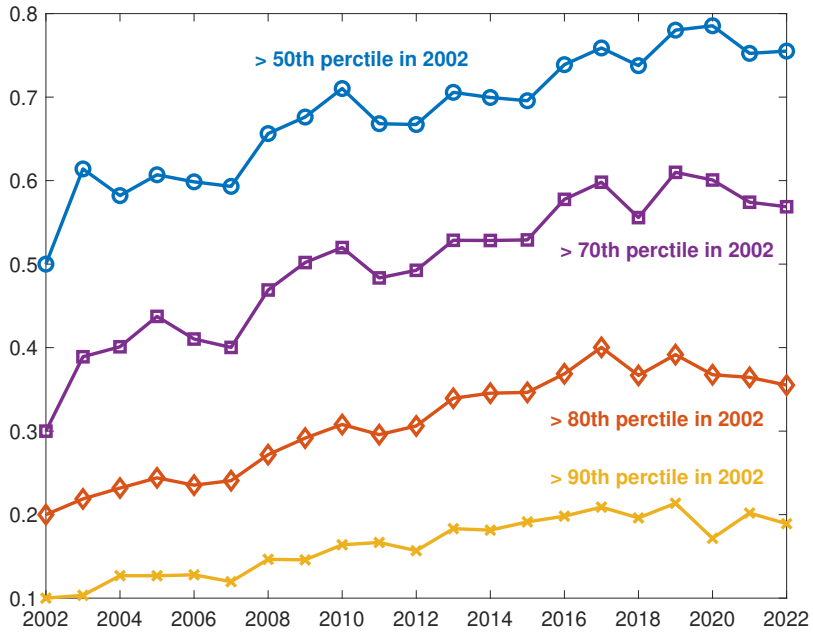
Notes: Estimates from the Duke CFO Survey. Column (1) shows estimates from a regression of the square of individual one-year-ahead real GDP growth errors on firm size (employment) and sector and time fixed effects. Firm size is measured discretely (values 1-5), depending on which quintile firm employment is in relative to the 2020-employment distribution. Column (2) controls for the familiarity of the respondent with the concept of GDP. Columns (3) and (4) consider the absolute value of individual errors. GDP errors are normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. Robust clustered standard errors in parentheses. Sample: 2020-2022.

Figure A.2: Time evolution of firm size



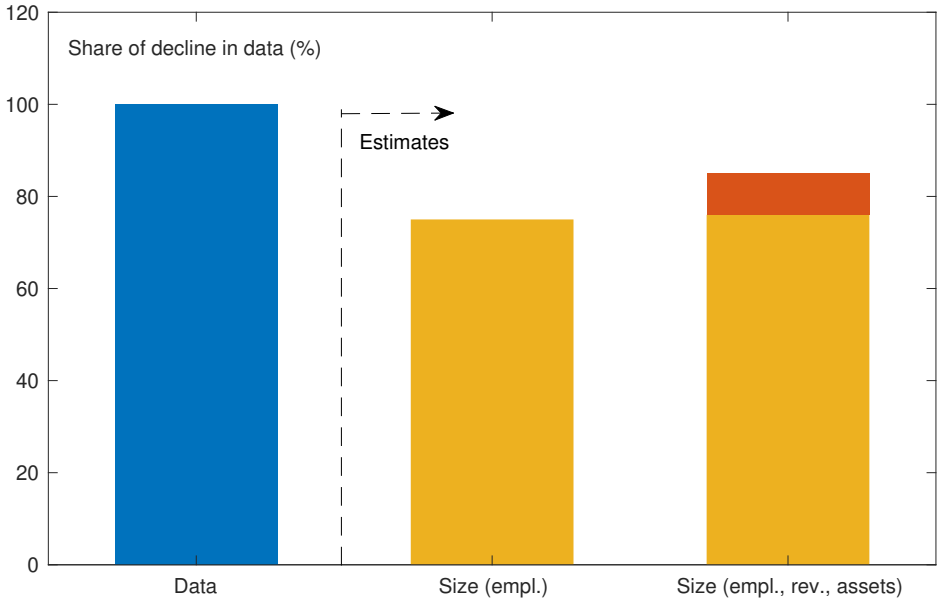
Note: Data from the I/B/E/S-Compustat sample. The panel shows the percentage increase from 2002 in the share of firms in a given year with employment exceeding the xth percentile of the 2002-employment distribution. The 50th, 80th, 90th percentile of the 2002-employment distribution correspond to around 1,000, 7,000, and 18,000 employees, respectively. Table A.3 in the Appendix shows the associated regression results.

Figure A.3: Time evolution of relative firm size



Note: Data from I/B/E/S-Compustat. The figure shows the share of firms in a given year with employment exceeding the xth percentile of the 2002-employment distribution. The 50th, 70th, 80th, 90th percentile of the 2002-employment distribution correspond to around 1,000, 7,000, and 18,000 employees, respectively.

Figure A.4: Size and accuracy simulation



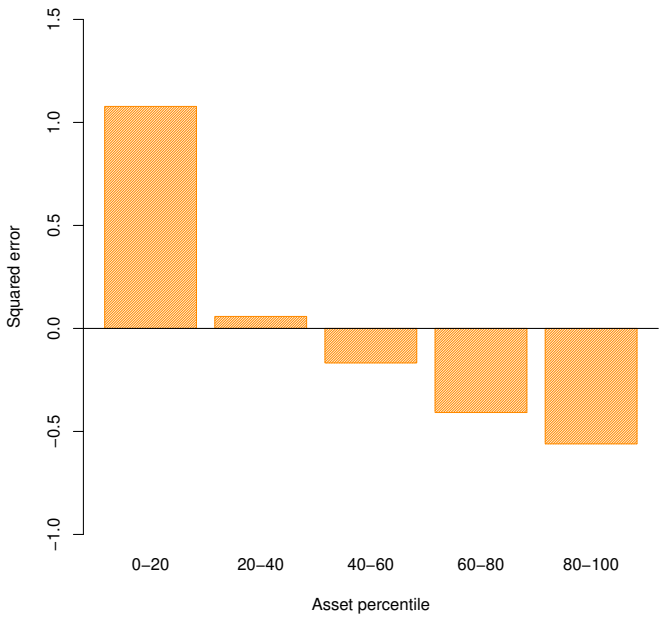
Note: The figure shows the overall rise in revenue accuracy in the I/B/E/S-Compustat sample (Figure 1) from its average value in 2002-2005 to 2022. The figure compares the decline to that implied from the change in the firm-size distribution, using the estimates in Table I Column 4 (with/without the inclusion of real firm assets and real firm revenue as further control variables). Firm real revenues (assets) are measured by the quintile the firm's revenues (assets) are in at time t relative to the 2002-revenue (asset) distribution. Firms revenues and assets are deflated by CPI-U from FRED.

Table A.6: Time evolution of asset size

	Panel (b): asset size and time*			
	50th perc.	70th perc.	80th perc.	90th perc.
	(1)	(2)	(3)	(4)
Time	0.013***	0.011***	0.007***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.408***	0.282***	0.164***	0.102***
	(0.016)	(0.011)	(0.005)	(0.005)
Observations	21	21	21	21
Residual std. error	0.031	0.026	0.016	0.016
F Statistic	126.291***	150.569***	145.456***	62.575***

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. The table estimates the coefficients of the share of firms with assets greater than the xth percentile of firms in 2002 on time. Assets are deflated by CPI-U from FRED. Robust standard errors in parentheses. The 50th, 70th, 80th, and 90th percentile of the 2002-asset distribution correspond to c. 288, 512, 992, and 3,728 million USD. Sample: 2002–2022.

Figure A.5: Revenue accuracy across the asset distribution



Note: The figure plots the difference between the average squared error of one-year-ahead log-revenue expectations from I/B/E/S-Compustat within size (asset) quintiles and the overall average taken across all size levels. Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Table A.7 reports the coefficient estimates, controlling for firm characteristics. Sample: 2002-2022.

Table A.7: Robustness of revenue expectations, firm size, and time relationship

	Absolute error		*Squared error*		*Squared error log*
	(1)	(2)	(3)	(4)	(5)
Firm size	-0.353^{***}	-0.293^{***}	-0.560^{***}	-0.450^{***}	
	(0.034)	(0.040)	(0.072)	(0.084)	
Firm assets					-0.243^{***}
					(0.056)
Time	-0.006		-0.003		
	(0.006)		(0.012)		
Firm age	0.004	0.023	0.046	0.058	0.022
	(0.016)	(0.024)	(0.035)	(0.052)	(0.031)
Rev. volatility		0.010^{***}		0.010^{**}	0.004
		(0.002		(0.002)	(0.011)
Observations	12,489	6,819	12,489	6,819	6,834
Time FE	×	✓	×	✓	✓
Sector FE	✓	✓	✓	✓	✓
F Statistic	10.460^{***}	7.704^{***}	5.494^{***}	5.043^{***}	2.083^{***}

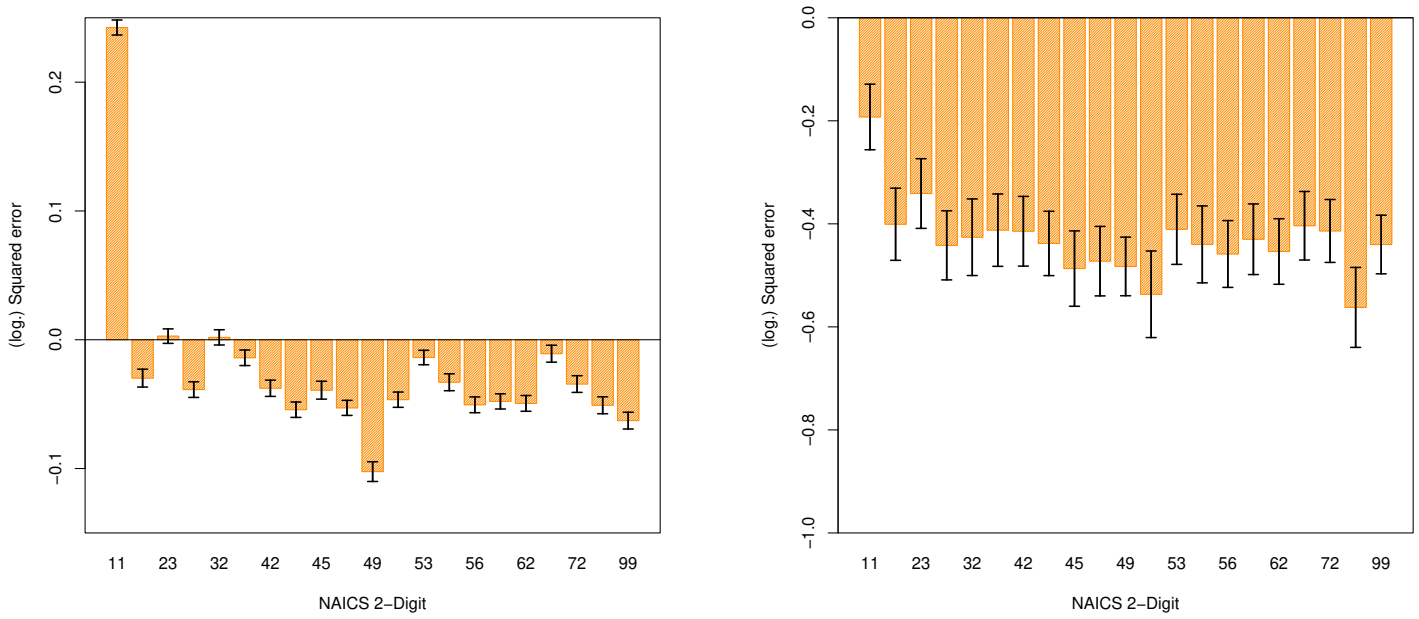
Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates from a regression of the absolute value of individual one-year-ahead revenue errors on firm size (employment), controlling for time, firm age, and sector fixed effects (NAICS-4). Firm size is measured based on which quintile the firm's employment level is at time t relative to the 2002-employment distribution. Column (2) considers the same regression specification but includes time fixed effects, as well as the rolling four-year volatility of revenue. Columns (3) and (4) consider the same specifications studied in Columns (1) and (2), but instead use the squared value of individual errors as the dependent variable. Finally, Column (5) uses the squared value of individual one-year-ahead revenue errors and measures firm size based on which quintile the firm's asset level is at time t relative to the asset distribution. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Table A.8: Revenue expectations, firm size, and fixed effects

	Squared log-revenue error		
	(1)	(2)	(3)
Firm size	-0.343^{***}	-0.438^{***}	-0.316^{***}
	(0.068)	(0.054)	(0.064)
Firm age	0.054^{*}		0.023
	(0.030)		(0.029)
Log. revenue volatility	-0.006		-0.006
	(0.011)		(0.011)
Observations	6,809	12,488	6,809
Sector FE	✓	×	×
Time FE	✓	×	×
Time×Sector FE	×	✓	✓
F statistic	2.441^{***}	2.401^{***}	1.438^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates of the squared value of one-year-ahead log-revenue errors on firm size (employment) and sector (NAICS-4) and time fixed effects. We also control for firm age and the individual four-year-rolling average of the volatility of revenue. Firm size is measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Columns (2) and (3) adds time×sector (NAICS-2) fixed effects. Revenue errors are scaled by firm capital and normalized by the overall average absolute error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Figure A.6: Sectoral heterogeneity and the time and size relationship



Panel (a): time relationship

Panel b: size relationship

Note: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the squared value of individual one-year-ahead (log-) revenue errors on time for different NAICS 2-digit sectors. Panel (b): estimate of the coefficient of the squared value of individual one-year-ahead (log-) revenue errors on firm size for different NAICS 2-digit sectors. Firm size is measured measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.9: Other variables (profits and capex)

	Panel (a): errors and time			
	Profits		Capex	
	Abs. error	Sqr. error	Abs. error	Sqr. error
Time	-0.039^{***}	-0.068^{***}	-0.012^{***}	0.001
	(0.010)	(0.024)	(0.003)	(0.011)
Constant	1.398^{***}	1.691^{***}	1.126^{***}	0.986^{***}
	(0.126)	(0.303)	(0.035)	(0.120)
Observations	2,487	2,487	1,839	1,839
Residual std. error	2.482	6.187	1.871	6.982
F statistic	15.27^{***}	7.385^{***}	12.91^{***}	0.011
	Panel (b): errors and size			
	Profits		Capex	
	Abs. error	Sqr. error	Abs. error	Sqr. error
Firm size	-0.361^{***}	-0.578^{***}	-0.074^{***}	-0.130^{***}
	(0.082)	(0.195)	(0.018)	(0.047)
Firm age	-0.029	0.047	-0.070^{***}	-0.117^{*}
	(0.080)	(0.147)	(0.016)	(0.061)
Constant	0.456	0.680	-0.113^{***}	-0.144
	(0.314)	(0.486)	(0.042)	(0.165)
Observations	2,487	2,487	1,839	1,839
Sector FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Residual std. error	2.347	6.209	1.740	6.744
F statistic	2.694^{***}	1.056	7.906^{***}	2.849^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead profit (capex) errors on time. Forecast errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead errors on firm size, controlling for firm age, and time and sector (NAICS level 4) fixed effects. Firm size is measured as in Table I. Robust (clustered) standard errors in parentheses. Sample: 2004–2022 (profits) and 2002–2022 (capex).

Table A.10: Percentage increase in revenue

	Absolute error		*Squared error*		
	(1)	(2)	(3)	(4)	(5)
Firm size	-0.210^{***}	-0.177^{***}	-0.372^{***}	-0.308^{***}	
	(0.025)	(0.031)	(0.056)	(0.064)	
Time	0.0005		0.005		-0.014^{**}
	(0.004)		(0.008)		(0.007)
Firm age	-0.050^{***}	0.009	-0.095^{**}	0.026	
	(0.016)	(0.025)	(0.039)	(0.051)	
Rev. volatility (pct.)		0.004^{***}		0.005^{***}	
		(0.001)		(0.002)	
Observations	10,083	5,700	10,083	5,700	10,138
Time FE	×	✓	×	✓	×
Sector FE	✓	✓	✓	✓	×
F Statistic	5.589^{***}	4.648^{***}	3.491^{***}	3.163^{***}	6.747^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates from a regression of the absolute value of individual one-year-ahead revenue growth errors (in pct.) on firm size (employment), controlling for time, firm age, and sector fixed effects (NAICS-4). Firm size is measured based on which quintile the firm's employment level is at time t relative to the 2002-employment distribution. Column (2) considers the same specification but includes time fixed effects, as well as the rolling four-year volatility of revenue. Columns (3) and (4) consider the same specifications studied in Columns (1) and (2), but instead use the squared value of individual errors as the dependent variable. Finally, Column (5) considers the raw correlation with time. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of forecast errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Table A.11: Large acquisitions and forecast accuracy

	Squared error	Absolute error
	(1)	(2)
Large acquisitions	−0.140*	−0.061*
	(0.007)	(0.033)
Large acquisitions(-1)	−0.119*	−0.082***
	(0.066)	(0.031)
Large acquisitions(-2)	−0.106*	−0.079***
	(0.062)	(0.031)
Large acquisitions(-3)	0.040	−0.040
	(0.068)	(0.031)
Firm age	0.038	0.005
	(0.034)	(0.017)
Log revenue volatility	0.027**	0.017***
	(0.013)	(0.005)
Observations	5,108	5,108
Sector FE	✓	✓
Time FE	✓	✓
Residual std. error	2.324	1.048
F statistic	1.569***	2.954***

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. The table estimates the coefficient of the squared value (absolute value) of individual one-year-ahead log-revenue errors on firm acquisitions, controlling for firm age, revenue volatility (rolling 4-year average), and time and sector (NAICS level 4) fixed effects. Errors are scaled by a firm's tangible capital stock and normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. A large acquisitions is defined as one above 5 percent of a firm's assets, consistent with the definition in Ottonello and Winberry (2020). Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.12: Accuracy and intangible capital

	Accuracy of expectations	
	Sqr. error	Abs. error
Firm acq. stock of intangibles	-0.070*	-0.038***
	(0.040)	(0.015)
Firm size	-0.405***	-0.211***
	(0.041)	(0.019)
Firm age	-0.051	-0.034***
	(0.034)	(0.013)
Observations	11,371	11,371
Sector FE	✓	✓
Time FE	✓	✓
F statistic	3.721***	6.499***

Notes: Panel least-squares estimates from I/B/E/S-Compustat. The table estimates the relationship between of the stock of acquired intangibles and the accuracy of firms' log-revenue expectations. The stock of acquired intangibles accounts adjusts amortization and take-outs financial goodwill (Compustat: `INTAN+AM-GDWL`), and the nominal stock is deflated. This is in accordance with Chiavari and Goraya (2023). Column (1) considers the squared value of individual errors, while Column (2) considers the absolute value. Errors are normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. All estimates controls for time and sector (NAICS-4) fixed effects. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.13: Accuracy and IDC data spend

	Accuracy of expectations		
	Data	Baseline	Rent-extracting
log(1 + data spend)	-0.670***	-0.566	-0.538
	(0.212)	(.)	(.)
Observations	177	.	.
Sector FE	✓	.	.
F statistic	4.897***	.	.
R2	0.33	.	.

Notes: Column 1: least-squares estimates from the merger of the I/B/E/S-Compustat sample and the IDC data set. The column estimates the coefficient of the squared value of individual one-year-ahead log-revenue errors on firm data spend (Section 6). Data spend equals the sum of hardware, software, and IaaS expenditures in the IDC. Estimates control for sector (NAICS-2) fixed effects. Robust (clustered) standard errors in parentheses. Sample: 2021. Column 2-3 estimate the analogous relationship in the calibrated model environments.

Table A.14: Accuracy and growth in firm s ize

	Employment		*Revenue*	
	(1)	(2)	(3)	(4)
Informed firms	0.100***	0.110***	0.087**	0.081**
	(0.037)	(0.037)	(0.037)	(0.039)
Initial employment	0.067***	0.066***		
	(0.003)	(0.003)		
Initial revenue			1.175***	1.152***
			(0.001)	(0.001)
Observations	10,186	10,186	10,234	10,234
Sector FE	✓	×	✓	×
Time FE	✓	✓	✓	✓
Industry FE	×	✓	×	✓
Added controls	age	age	age	age
F statistic	1,681***	272.7***	1,573***	271.2***

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Columns (1) and (2) report estimates of a firm's subsequent employment (2007-2022) on whether a firm was "informed" or not in the initial period (2002-2007), the firm's initial age, as well as the firm's initial employment (2002). We further control for time and sector (NAICS-2) or industry (NAICS-4) fixed effects. Columns (3) and (4) report estimates which instead focus on a firm's revenue. Robust (clustered) standard errors in parentheses. Sample: 2002-2022

B Proofs and Derivations for Sections 3 and 4

Consumer Tastes. Here, we provide a simple microfoundation for the formulation of household preferences in (1). Suppose that instead the utility of household i is given by:

$$\mathcal{U}_i = C_i = \left[\int_0^1 \left(\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}}\right)^{\frac{\theta-1}{\theta}} \cdot dj\right]^{\frac{\theta}{\theta-1}}, \tag{A1}$$

where c_{ij}^{r} and c_{ij}^{b} are the consumptions of the red- and blue-type variety j, respectively, and where δ_{ij}^{r} and δ_{ij}^{b} are taste shocks that have the following mapping to the demand state ω_j:

$$\delta_{ij}^{\mathrm{r}} = \begin{cases} \varsigma_{ij} \cdot \bar{\sigma} & \text{if } \omega_j = \text{red} \\ \varsigma_{ij} \cdot \underline{\sigma} & \text{otherwise} \end{cases} \text{ and } \delta_{ij}^{\mathrm{b}} = \begin{cases} \varsigma_{ij} \cdot \bar{\sigma} & \text{if } \omega_j = \text{blue} \\ \varsigma_{ij} \cdot \underline{\sigma} & \text{otherwise} \end{cases}. \tag{A2}$$

Given this formulation, if the demand state is ω_j and if the variety j supplied in the market in equilibrium is of type $x_j = \omega_j$, then optimal consumption choice implies:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \varsigma_{ij} \cdot \bar{\sigma} \cdot c_{ij}, \tag{A3}$$

where c_{ij} are the total units of the variety consumed by household i. If instead the variety j available in the market is of type $x_j \neq \omega_j$, then:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \varsigma_{ij} \cdot \underline{\sigma} \cdot c_{ij}. \tag{A4}$$

Therefore, given the types $\{x_j\}_j$ of varieties chosen by firms in equilibrium, this formulation of household preferences coincides with that in Equation (1) in the main text. □

Proof of Proposition 1. Using household demand in Equation (10) and market clearing for variety j, for given choices of inputs and information, (x_j, n_j, ι_j), firm j's expected profits equal:

$$\mathbb{E}\left[\pi_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[\left(\delta_j \cdot A_j\right)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot \left(n_j + \chi \cdot \iota_j\right), \tag{A5}$$

where δ_j is given by Equation (10), and where due to independence of the shocks υ_j and ω_j:

$$\mathbb{E}\left[\left(\delta_j \cdot A_j\right)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}} | s_j^{\omega}, \tau_j^{\omega}\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}} | \mu_j, s_j^{\upsilon}, \tau_j^{\upsilon}\right]. \tag{A6}$$

where:

$$\mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}}|s_j^\omega,\tau_j^\omega\right]=\mathbb{P}\left(\omega_j=x_j|s_j^\omega,\tau_j^\omega\right)\cdot\delta\left(\bar{\sigma}\right)^{\frac{\theta-1}{\theta}}+\left(1-\mathbb{P}\left(\omega_j=x_j|s_j^\omega,\tau_j^\omega\right)\right)\cdot\delta\left(\underline{\sigma}\right)^{\frac{\theta-1}{\theta}}. \quad \text{(A7)}$$

Now, because the demand shifter of the "representative consumer" is strictly higher when the firm customizes the variety to the consumer's tastes, i.e., $\delta\left(\bar{\sigma}\right)>\delta\left(\underline{\sigma}\right)$, and because signals are weakly informative, firm j optimally chooses to set:

$$x_j=s_j^\omega\in\{\text{red},\text{blue}\} \quad\Longrightarrow\quad \mathbb{P}\left(\omega_j=x_j|s_j^\omega,\tau_j^\omega\right)=\tau_j^\omega. \quad \text{(A8)}$$

Maximizing expected profits in Equation (A5) with respect to n_j now yields:

$$n_j=\mathbb{E}\left[\left(\delta_j\cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]^\theta\cdot\Omega, \quad \text{(A9)}$$

as stated in the proposition. Lastly, plugging back the optimal choices of x_j and n_j into the expression for expected profits in Equation (A5), we get:

$$\mathbb{E}\left[\pi_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]=\left(\frac{1}{\theta-1}\cdot\mathbb{E}\left[n_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]-\chi\cdot\iota_j\right)\cdot w. \quad \text{(A10)}$$

Thus, at *Stage 1*, the firm produces information if and only if:

$$\frac{1}{\theta-1}\cdot\left(\mathbb{E}\left[n_j|\mu_j,\bar{\boldsymbol{\tau}}_j\right]-\mathbb{E}\left[n_j|\mu_j,\underline{\boldsymbol{\tau}}_j\right]\right)\geq\chi. \quad \text{(A11)}$$

Finally, to arrive at the expression for the expected employment in the text, observe that:

$$\mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}}|s_j^\omega,\tau_j^\omega\right]=\tau_j^\omega\cdot\delta\left(\bar{\sigma}\right)^{\frac{\theta-1}{\theta}}+\left(1-\tau_j^\omega\right)\cdot\delta\left(\underline{\sigma}\right)^{\frac{\theta-1}{\theta}}, \quad \text{(A12)}$$

which is independent of the realization of s_j^ω, and:

$$\mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j,s_j^v,\tau_j^v\right]=\exp^{\frac{\theta-1}{\theta}\cdot\mu_j}\cdot\exp^{\frac{\theta-1}{\theta}\cdot\frac{\tau_j^v}{\tau_j^v+\tau_a}\cdot s_j^v+\frac{1}{2}\cdot\left(\frac{\theta-1}{\theta}\right)^2\cdot\frac{1}{\tau_j^v+\tau_a}}, \quad \text{(A13)}$$

which implies that:

$$\mathbb{E}\left[\mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j,s_j^v,\tau_j^v\right]^\theta|\mu_j,\tau_j^v\right]=\exp^{(\theta-1)\cdot\mu_j}\cdot\exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\theta}\cdot\frac{\tau_a+\theta\cdot\tau_j^v}{\tau_a+\tau_j^v}\cdot\frac{1}{\tau_a}}. \quad \text{(A14)}$$

Therefore, the expected employment of firm j, conditional on its information choice is:

$$\begin{aligned}\mathbb{E}\left[n_j|\mu_j,\boldsymbol{\tau}_j\right] &= \left[\tau_j^{\omega}\cdot\delta\left(\bar{\sigma}\right)^{\frac{\theta-1}{\theta}}+\left(1-\tau_j^{\omega}\right)\cdot\delta\left(\underline{\sigma}\right)^{\frac{\theta-1}{\theta}}\right]^{\theta}\cdot\exp^{(\theta-1)\cdot\mu_j}\cdot\exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\theta}\cdot\frac{\tau_a+\theta\cdot\tau_j^{\upsilon}}{\tau_a+\tau_j^{\upsilon}}\cdot\frac{1}{\tau_a}}\cdot\Omega \\ &= \exp^{(\theta-1)\cdot\mu_j}\cdot g\left(\boldsymbol{\tau}_j\right)^{\theta-1}\cdot\Omega, \end{aligned} \tag{A15}$$

where $g(\cdot)$ is as defined in Equation (16). □

Proof of Corollary 1. The proof follows immediately from the definition of tfp_j in Equation (20), the fact that firm optimality implies that, conditional on $(\mu_j,\boldsymbol{\tau}_j)$, the demand shifter δ_j (as defined by Equation (10)) equals $\delta(\bar{\sigma})$ with probability τ_j^{ω} and $\delta(\underline{\sigma})$ with probability $1-\tau_j^{\omega}$, combined with the fact that $\log(A_j)|\mu_j\sim\mathcal{N}(\mu_j,\tau_a^{-1})$. □

Proof of Corollary 2. From the optimality condition for n_j in Equation (13) and the definition of mrp_j in Equation (23), we have:

$$\text{mrp}_j = \log\left(p_j\cdot y_j\right)-\log\left(\mathbb{E}\left[p_j\cdot y_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]\right)-\log\left(\frac{\theta}{\theta-1}\cdot w\right). \tag{A16}$$

It therefore follows immediately that:

$$\mathbb{VAR}\left[\text{mrp}_j|\mu_j,\boldsymbol{\tau}_j\right] = \mathbb{VAR}\left[\text{error}_j|\mu_j,\boldsymbol{\tau}_j\right]. \tag{A17}$$

Next, using household demand and market clearing for variety j, we have that firm j's revenues are given by the expression, as also used in the proof of Proposition 1:

$$p_j\cdot y_j = \left(\delta_j\cdot A_j\right)^{\frac{\theta-1}{\theta}}\cdot n_j^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}}. \tag{A18}$$

Thus, we have that:

$$\begin{aligned}\text{error}_j &= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\cdot A_j\right)-\mathbb{E}\left[\log\left(\delta_j\cdot A_j\right)|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]\right) \\ &= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\right)-\mathbb{E}\left[\log\left(\delta_j\right)|\tau_j^{\omega}\right]\right)+\frac{\theta-1}{\theta}\cdot\left(\log\left(A_j\right)-\mathbb{E}\left[\log\left(A_j\right)|\mu_j,s_j^{\upsilon},\tau_j^{\upsilon}\right]\right) \\ &= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\right)-\mathbb{E}\left[\log\left(\delta_j\right)|\tau_j^{\omega}\right]\right)+\frac{\theta-1}{\theta}\cdot\left(\frac{\tau^a}{\tau_j^{\upsilon}+\tau_a}\cdot\upsilon_j-\frac{\tau_j^{\upsilon}}{\tau_j^{\upsilon}+\tau_a}\cdot\varepsilon_j\right). \end{aligned} \tag{A19}$$

Therefore:

$$\mathbb{VAR}\left[\text{mrp}_j|\mu_j,\boldsymbol{\tau}_j\right] = \left(\frac{\theta-1}{\theta}\right)^2 \cdot \mathbb{VAR}\left[\log\left(\delta_j\right)|\tau_j^{\omega}\right] + \left(\frac{\theta-1}{\theta}\right)^2 \cdot \frac{1}{\tau_j^{v}+\tau_a} \tag{A20}$$

$$= \mathbb{VAR}\left[\text{tfp}_j|\mu_j,\boldsymbol{\tau}_j\right] + \left(\frac{\theta-1}{\theta}\right)^2 \cdot \left(\frac{1}{\tau_j^{v}+\tau_a} - \frac{1}{\tau_a}\right). \tag{A21}$$

□

Proof of Corollary 3. The expression for the marginal-type $\bar{\mu}$ that is just indifferent to producing information follows from the optimality condition for information production in Equation (18) and the expression for a firm's expected employment in Equation (19). Next, consider two firms, j and j', with mean productivities $\mu_j < \bar{\mu} < \mu_{j'}$, so that firm j' produces information while firm j does not.

First, the statement that firm j' on average has a higher and less dispersed tfp_j follows from Corollary 1 and the fact that $\mu_{j'} > \mu_j$ and $\tau_{j'}^{\omega} > \tau_j^{\omega}$.

Second, the statement that firm j' has less dispersed mrp_j follows from Corollary 2 and the fact that $\mu_{j'} > \mu_j$ and $\tau_{j'}^{v} > \tau_j^{v}$, combined with the fact that it has less dispersed tfp_j.

Finally, that $\mathbb{E}[n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[n_j|\mu_j,\boldsymbol{\tau}_j]$ follows from the expression for a firm's expected size in Equation (19), the definition of the information shifter in Equation (16), combined with the fact that $\mu_{j'} > \mu_j$ and $\boldsymbol{\tau}_{j'} > \boldsymbol{\tau}_j$. Clearly, it then also follows that (i) $\mathbb{E}[n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}]+\chi\cdot\iota_{j'} > \mathbb{E}[n_j|\mu_j,\boldsymbol{\tau}_j]+\chi\cdot\iota_j$, since $\iota_{j'} = 1 > 0 = \iota_j$; and (ii) $\mathbb{E}[p_{j'}\cdot y_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[p_j\cdot y_j|\mu_j,\boldsymbol{\tau}_j]$ and $\mathbb{E}[p_{j'}\cdot y_{j'} - w\cdot n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[p_j\cdot y_j - w\cdot n_j|\mu_j,\boldsymbol{\tau}_j]$, since both are proportional to a firm's expected employment (Equation (19)). Lastly, it follows that expected profits are higher for j', since firm j' has chosen to produce information. □

Proof of Lemma 1. Equation (29) in the text follows by using the definition of aggregate TFP, $\mathcal{A}$, and the definition of aggregate employment in goods production, $\mathcal{N}$, combined with the optimality conditions in Proposition 1. Moreover:

$$\int_0^1 \mathbb{E}\left[\left(\delta_j\cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]^{\theta}\cdot dj = \int_0^1 \exp^{(\theta-1)\cdot\mu_j}\cdot g\left(\boldsymbol{\tau}_j\right)^{\theta-1}\cdot dj$$

$$= g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1}\cdot\int_{-\infty}^{\bar{\mu}}\exp^{(\theta-1)\cdot\mu}\cdot d\Phi\left(-\mu\cdot\sqrt{\tau_\mu}\right) + g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1}\cdot\int_{\bar{\mu}}^{\infty}\exp^{(\theta-1)\cdot\mu}\cdot d\Phi\left(-\mu\cdot\sqrt{\tau_\mu}\right), \tag{A22}$$

where, using the definition $\xi\left(\bar{\mu}\right) \equiv \Phi\left(-\bar{\mu}\cdot\sqrt{\tau_\mu} + \frac{\theta-1}{\sqrt{\tau_\mu}}\right)$, we have:

$$\int_{\bar{\mu}}^{\infty}\exp^{(\theta-1)\cdot\mu}\cdot d\Phi\left(-\mu\cdot\sqrt{\tau_\mu}\right) = \int_{\bar{\mu}}^{\infty}\frac{1}{\sqrt{2\cdot\pi\cdot\tau_\mu^{-1}}}\cdot\exp^{-\mu^2\cdot\frac{1}{2}\cdot\tau_\mu+(\theta-1)\cdot\mu}\cdot d\mu \quad = \exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\tau_\mu}}\cdot\xi\left(\bar{\mu}\right),$$

and, similarly:

$$\int_{-\infty}^{\bar{\mu}} \exp^{(\theta-1)\cdot\mu} \cdot d\Phi\left(-\mu \cdot \sqrt{\tau_\mu}\right) = \exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\tau_\mu}} \cdot \left(1-\xi\left(\bar{\mu}\right)\right). \tag{A23}$$

The result then follows by replacing these expressions into Equation (29). □

Proof of Proposition 2. That an equilibrium exists and is unique follows from the observation that: (i) the map $\bar{\mu}: \mathbb{R}^+ \to \mathbb{R}$ defined by Equation (27) is continuous, decreasing, and goes from ∞ to $-\infty$ as Ω ranges from 0 to ∞, and (ii) the map $\Omega: \mathbb{R} \mapsto \mathbb{R}^+$ defined by Equation (32) is continuous, increasing, and goes from $\frac{N-\chi}{\lim_{\bar{\mu}\to-\infty}\mathcal{A}(\bar{\mu})^{\theta-1}}$ to $\frac{N}{\lim_{\bar{\mu}\to\infty}\mathcal{A}(\bar{\mu})^{\theta-1}}$ as $\bar{\mu}$ ranges from $-\infty$ to ∞. The intersection of these maps yields the equilibrium $\mu^\star$; the expressions for TFP, consumption and welfare follow immediately from Lemma 1. □

Proof of Proposition 3. Combining Equations (27) and (32), together with the expression for aggregate TFP in Lemma 1, we have that the equilibrium $\mu^\star$ satisfies:

$$\exp^{(\theta-1)\cdot\mu^\star} = \exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\tau_\mu}} \cdot (\theta-1) \cdot \chi \cdot \frac{\frac{g(\underline{\boldsymbol{\tau}})^{\theta-1}}{g(\bar{\boldsymbol{\tau}})^{\theta-1}-g(\underline{\boldsymbol{\tau}})^{\theta-1}} + \xi\left(\mu^\star\right)}{N-\chi\cdot\Phi\left(-\mu^\star\cdot\sqrt{\tau_\mu}\right)}. \tag{A24}$$

The left-hand side of Equation (A24) is monotonically increasing in $\mu^\star$, whereas the right-hand side is monotonically decreasing in $\mu^\star$. As a result, either a decline in χ or a rise in $\bar{\boldsymbol{\tau}}$, both of which imply a decline in the right-hand side, lead to a decline in $\mu^\star$ and thus a rise in the share of firms that produce information. From Lemma 1, it then follows that aggregate TFP rises as well; in the case of a rise in $\bar{\boldsymbol{\tau}}$ it rises both *directly* as $g\left(\bar{\boldsymbol{\tau}}\right)$ increases and *indirectly* through the fall in $\mu^\star$. As for aggregate consumption and welfare, it is straightforward to show that Equation (A24) implies, after some algebra, that:

$$\frac{d}{d\bar{\mu}} \cdot \left(\mathcal{A}\left(\bar{\mu}, g\right) \cdot \left[N-\chi\cdot\Phi\left(-\bar{\mu}\cdot\sqrt{\tau_\mu}\right)\right]\right)|_{\bar{\mu}=\mu^\star} = 0. \tag{A25}$$

Hence, by an envelope argument, we need only consider the direct effect of changes in χ or $\bar{\boldsymbol{\tau}}$ on $C^\star$. But, these are clearly positive, since, for a given $\mu^\star$, a lower χ reduces costs of information production, whereas a higher $\bar{\boldsymbol{\tau}}$ raises aggregate TFP. □

C Proofs and Derivations for Section 5

C.1 Optimal Trading Mechanism

We provide detailed derivations for the optimal trading mechanism. Since at *Stage* 3, the firm's supply of variety j is inelastic and given by $A_j \cdot n_j$, the firm's objective is to maximize

its revenues (Equation (35)) subject to incentive compatibility (Equation (36)), individual rationality (Equation (37)), and feasibility (Equation (38)) constraints.

We conjecture (and then verify) that the individual rationality (IR) constraint of the h-type and the incentive compatibility (IC) constraint of the ℓ-type are slack in each segment. Given this, it is optimal for the firm to set $t_j(h|s)$ and $t_j(\ell|s)$ so that both the IC-constraint of the h-type and the IR-constraint of the ℓ-type in each segment s bind. Moreover, it is clear that the firm will allocate all its goods to consumers, so that the feasibility constraint binds as well. As a result, the problem reduces to choosing the quantities $\{q_j(\varsigma|s)\}$ to maximize:

$$\sum_{s\in\{h,\ell\}} \mathbb{P}(s_{ij}=s)\cdot\Big[\mathbb{P}(\varsigma_{ij}=h|s_{ij}=s)\cdot h^{\frac{\theta-1}{\theta}}\cdot q_j(h|s)^{\frac{\theta-1}{\theta}} + \left(\ell^{\frac{\theta-1}{\theta}} - \mathbb{P}(\varsigma_{ij}=h|s_{ij}=s)\cdot h^{\frac{\theta-1}{\theta}}\right)\cdot q_j(\ell|s)^{\frac{\theta-1}{\theta}}$$

(A26)

subject to the feasibility constraint in Equation (38) holding with equality. The solution to this problem yields the allocations given by Equations (39) and (40).

Next, denote the optimal allocation of the type-ς consumer in segment s by $\left(t_j^\star(\varsigma|s), q_j^\star(\varsigma|s)\right)$, and let us verify the conjecture that the IC-constraint of the ℓ-type and the IR-constraint of the h-type are satisfied at the optimum. The latter follows by combining the binding IC-constraint of the h-type and the binding IR-constraint of the ℓ-type:

$$\begin{aligned} t_j^\star(h|s) &= \left(\varsigma_h\cdot\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}} - \left(\varsigma_h^{\frac{\theta-1}{\theta}} - \varsigma_\ell^{\frac{\theta-1}{\theta}}\right)\cdot\left(\sigma_j\cdot q_j^\star(\ell|s)\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}} \\ &\le \left(\varsigma_h\cdot\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}}. \end{aligned} \tag{A27}$$

The former follows from the fact that $q_j^\star(\ell|s)\le q_j^\star(h|s)$ and, again, by combining the binding IC-constraint of the h-type and the binding IR-constraint of the ℓ-type:

$$\begin{aligned} \left(\varsigma_\ell\cdot\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}} - t_j^\star(h|s) &= \left(\varsigma_\ell\cdot\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}} + \\ &\quad - \left[\left(\varsigma_h\cdot\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}} - \left(\varsigma_h^{\frac{\theta-1}{\theta}} - \varsigma_\ell^{\frac{\theta-1}{\theta}}\right)\cdot\left(\sigma_j\cdot q_j^\star(\ell|s)\right)^{\frac{\theta-1}{\theta}}\right]\cdot C^{\frac{1}{\theta}} \\ &= \left(\varsigma_\ell^{\frac{\theta-1}{\theta}} - \varsigma_h^{\frac{\theta-1}{\theta}}\right)\cdot\left[\left(\sigma_j\cdot q_j^\star(h|s)\right)^{\frac{\theta-1}{\theta}} - \left(\sigma_j\cdot q_j^\star(\ell|s)\right)^{\frac{\theta-1}{\theta}}\right]\cdot C^{\frac{1}{\theta}} \le 0, \end{aligned} \tag{A28}$$

and the fact that the ℓ-type earns zero surplus at the optimal allocation.

We now derive the expression for the ς-type consumer's surplus earned from a given allocation (t_{ij}, q_{ij}) offered by the mechanism, which we had assumed in the above derivations. To this end, note that the utility that consumer i gains from consuming q_{ij} units of variety j,

when her overall consumption is C_i, is given by:

$$\widetilde{u}_{ij}\left(q_{ij}\right)=\frac{\theta}{\theta-1}\cdot\left(\varsigma_{ij}\cdot\sigma_j\cdot q_{ij}\right)^{\frac{\theta-1}{\theta}}\cdot C_i^{\frac{1}{\theta}}. \tag{A29}$$

Suppose that λ_i is the (nominal) marginal value of income to the consumer. Then, the (nominal) surplus of the consumer is $\frac{\widetilde{u}_{ij}(q_{ij})}{\lambda_i}$. Since households are ex-ante identical, in any symmetric equilibrium it must be that $\lambda_i=\lambda$ and $C_i=C$ for all i. In equilibrium, therefore, market clearing implies that $\lambda=P^{-1}\cdot\frac{\theta}{\theta-1}\cdot\frac{\int(\delta_j^R\cdot A_j\cdot n_j)^{\frac{\theta-1}{\theta}}\cdot dj}{\int(\delta_j^S\cdot A_j\cdot n_j)^{\frac{\theta-1}{\theta}}\cdot dj}$, where P is the ideal price index and where the demand shifters δ_j^R and δ_j^S are defined by Equations (43) and (45). We normalize $P=\frac{\int(\delta_j^R\cdot A_j\cdot n_j)^{\frac{\theta-1}{\theta}}\cdot dj}{\int(\delta_j^S\cdot A_j\cdot n_j)^{\frac{\theta-1}{\theta}}\cdot dj}$ without loss of generality. Because in equilibrium all agents have correct expectations about λ and C, they understand that consumer i's willingness to pay for q_{ij} units of its variety is $\left(\varsigma_{ij}\cdot\sigma_j\cdot q_{ij}\right)^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}}$, as assumed in the main text.

Now, in the above, we assumed that the representative household's marginal value of income, λ, is well defined. For this to be the case, we need to be able to assign a utility value to an additional unit of income in the hands of the household. However, if firms were only offering the bundles $\left\{\left(t_j^\star(\varsigma|s),q_j^\star(\varsigma|s)\right)\right\}_{\varsigma,s}$, which are all accepted in equilibrium, the household would have no way to spend this additional income. To address this issue, we follow an approach similar to Bornstein and Peter (2024), and suppose that firms post additional latent allocations in their menus, which are not accepted in equilibrium but ensure that the consumer faces linear prices off-equilibrium. Thus, suppose that, in addition to the equilibrium bundles, each firm j offers to sell to the consumer any quantity q above $q_j^\star(h|h)$ for a total payment of:

$$t_j\left(q\right)=t_j^\star(h|h)+\kappa_j^\star\cdot\left(q-q_j^\star(h|h)\right)\ \text{ with }\ \kappa_j^\star=\left(\varsigma_h\cdot\sigma_j\cdot q_j^\star(h|h)\right)^{-\frac{1}{\theta}}\cdot C^{\frac{1}{\theta}}. \tag{A30}$$

By construction, given the marginal value of income $\lambda=\frac{\theta}{\theta-1}$, the household is indifferent to picking up an additional unit of the variety j if she is an h-type in segment h; otherwise, she prefers not to trade. Moreover, the marginal value of income to the household is now well defined since she can now spend the additional income (if she had any) to obtain units above $q_j^\star(h|h)$ for those varieties for which she is an h-type and is put by the firm into segment h.

C.2 Alternative Trading Environment: Indivisible Goods

In the main text, price discrimination arises through optimal menu pricing for divisible goods. In this Appendix, we show that the key objects of the model—the revenue-based demand shifter δ_j^R, the surplus-based demand shifter δ_j^S, and the wedge $\Psi_j=\delta_j^S/\delta_j^R$—also arise in a commonly studied environment with *indivisible goods and uniform pricing*.

Specifically, firms sell discrete units but choose the *quality* of each unit. Price discrimination occurs through prices that are contingent on the firm's signals about consumer tastes. These signal-contingent prices determine which consumers are served in each segment. Despite these microeconomic differences, the equilibrium objects that summarize firm behavior and the efficiency-rent extraction tradeoff are near identical to those in the main framework.

We modify the setting in Section 5 as follows. First, we assume that consumers have unit demand for each variety, but derive utility from the quality of that variety. Thus, c_{ij} now denotes the quality of variety j consumed by consumer i (see Equation (1)). Second, the firm's production function is

$$m_j \cdot q_j = A_j \cdot n_j, \tag{A31}$$

where m_j denotes the number of units produced and q_j their quality. Because effective output $A_j \cdot n_j$ is fixed at the production stage—once employment has been chosen and productivity has been realized—producing more units necessarily lowers the quality of each unit.

After producing m_j units of quality q_j, firm j must decide how to price these goods, given its signals $\{s_{ij}^{\varsigma}\}_i$ about consumer-specific tastes. Since consumers have unit demand, the firm effectively chooses between two pricing policies for each signal realization s_{ij}^{ς}: it can charge the willingness to pay of the low-demand type and serve both types, or charge the willingness to pay of the high-demand type and serve only high-demand consumers. Formally, the firm sets $p(s_{ij}^{\varsigma}) \in \{v_j(\ell),\, v_j(h)\}$, where:

$$v_j(\varsigma) \equiv (\varsigma \cdot \sigma_j \cdot q_j)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}. \tag{A32}$$

Let $\phi_j(s) \in \{0, 1\}$ denote the firm's pricing decision after observing signal s. If $\phi_j(s) = 1$, the firm posts the low price $v_j(\ell)$ and trades with both consumer types. If $\phi_j(s) = 0$, it posts the high price $v_j(h)$ and trades only with the high-demand type. Then the firm's optimal strategy is:

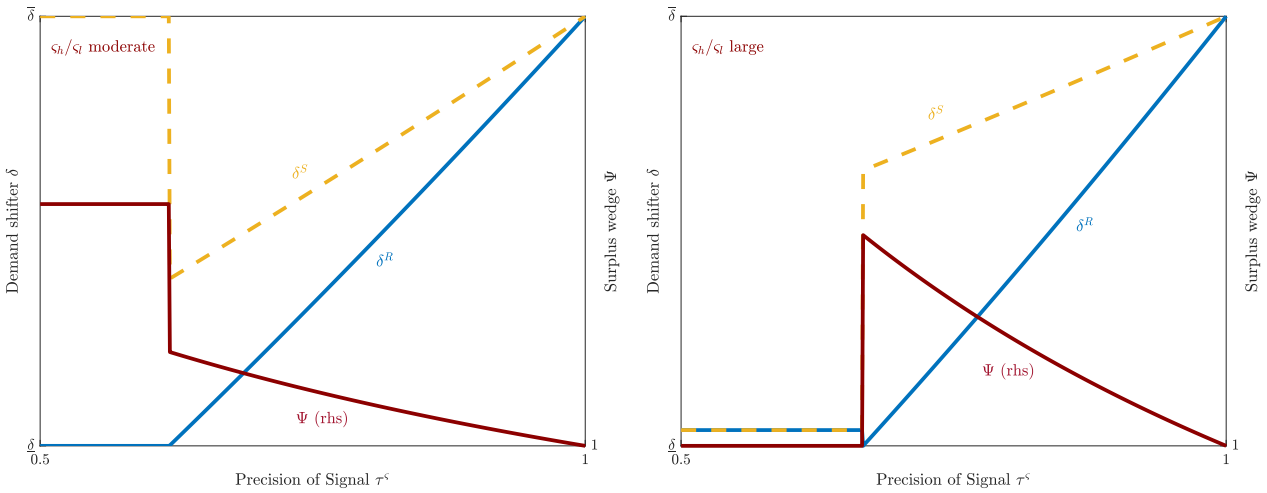
$$\phi_j(h) = \begin{cases} 1 & \text{if } 1 - \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^{\varsigma} = h\right) \cdot \left(\frac{h}{\ell}\right)^{\frac{\theta-1}{\theta}} \geq \frac{\theta-1}{\theta} \\ 0 & \text{otherwise} \end{cases}, \tag{A33}$$

$$\phi_j(\ell) = \begin{cases} 1 & \text{if } 1 - \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^{\varsigma} = \ell\right) \cdot \left(\frac{h}{\ell}\right)^{\frac{\theta-1}{\theta}} \geq \frac{\theta-1}{\theta} \cdot \frac{1}{\gamma + (1-\gamma) \cdot \mathbb{P}(s_{ij}^{\varsigma} = \ell | \varsigma_{ij} = \ell)} \\ 0 & \text{otherwise} \end{cases}, \tag{A34}$$

and the associated optimal quantity produced is:

$$m_j = \sum_s \mathbb{P}\left(s_{ij}^{\varsigma} = s\right) \cdot \left[\phi_j(s) + (1 - \phi_j(s)) \cdot \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^{\varsigma} = s\right)\right]. \tag{A35}$$

Figure C.1: Revenue- vs Surplus-Based Demand Shifters — Alternative Environment



(a) Moderate heterogeneity (b) Large heterogeneity

Note: The figure depicts the demand shifters δ_j^S and δ_j^R, and the wedge Ψ_j as a function of τ_j^ς. Panel (a) plots the case in which $\varsigma_h/\varsigma_\ell$ is relatively small, while Panel (b) plots the case in which $\varsigma_h/\varsigma_\ell$ is relatively high.

Thus, the firm trades off the revenue gain from serving additional consumers against the quality reduction required to do so. Because total effective output is pinned down by production, selling more units necessarily reduces the quality of each unit.

As in Section 5.3, all firm-level outcomes can be summarized by two demand shifters: a revenue-based shifter δ_j^R and a surplus-based shifter δ_j^S. These objects aggregate firm revenues and household utility generated by the optimal pricing strategy described above:

$$\delta_j^R(\tau_j^\varsigma, \sigma_j) \equiv \frac{\left(\sum_s \mathbb{P}\left(s_{ij}^\varsigma = s\right) \cdot \left[\phi_j(s) \cdot \ell^{\frac{\theta-1}{\theta}} + (1-\phi_j(s)) \cdot \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^\varsigma = s\right) \cdot h^{\frac{\theta-1}{\theta}}\right]\right)^{\frac{\theta}{\theta-1}}}{m_j} \cdot \sigma_j, \tag{A36}$$

and

$$\delta_j^S(\tau_j^\varsigma, \sigma_j) \equiv \Psi(\tau_j^\varsigma) \cdot \delta_j^R(\tau_j^\varsigma, \sigma_j), \tag{A37}$$

where the firm-level wedge is:

$$\Psi(\tau_j^\varsigma) \equiv \left(1 + \frac{\sum_s \mathbb{P}\left(s_{ij}^\varsigma = s\right) \cdot \phi_j(s) \cdot \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^\varsigma = s\right) \cdot \left(h^{\frac{\theta-1}{\theta}} - \ell^{\frac{\theta-1}{\theta}}\right)}{\sum_s \mathbb{P}\left(s_{ij}^\varsigma = s\right) \cdot \left[\phi_j(s) \cdot \ell^{\frac{\theta-1}{\theta}} + (1-\phi_j(s)) \cdot \mathbb{P}\left(\varsigma_{ij} = h | s_{ij}^\varsigma = s\right) \cdot h^{\frac{\theta-1}{\theta}}\right]}\right)^{\frac{\theta}{\theta-1}}. \tag{A38}$$

The key properties of these objects are identical to those established in Lemma 2. In particular, the revenue-based demand shifter δ_j^R is monotonically increasing in information precision τ_j^ς and converges to the surplus-based shifter δ_j^S as $\tau_j^\varsigma \to 1$. However, the surplus-

based shifter and the wedge Ψ_j may be non-monotonic in τ_j^S.

An increase in τ_j^S has two opposing effects. In the high-signal segment h, greater information allows the firm to exclude low-demand consumers by charging the high price, thereby increasing allocative distortions. In the low-signal segment ℓ, however, greater information allows the firm to charge the low price when appropriate and serve additional consumers, thereby reducing distortions.

When taste heterogeneity is moderate (Panel (a) of Figure C.1), distortions are initially small and the first effect dominates: improved information raises distortions and reduces surplus, generating the same conflict between efficiency and rent extraction emphasized in the main text. When heterogeneity is large (Panel (b)), distortions are initially severe and the second effect dominates, so improved information increases both surplus and revenue.

Finally, because all firm-level outcomes can be summarized by the same demand shifters (δ_j^R, δ_j^S), the equilibrium analysis of the rent-extracting economy in this alternative environment proceeds exactly as in Sections 5.3-5.5.

C.3 Proofs and Other Derivations

Proof of Lemma 2. See text. □

Proof of Proposition 4. The result follows immediately from the observation that the ex-post profits of firm j in the rent-extracting economy are given by Equation (42). □

Proof of Proposition 5. Given the equilibrium $\mu^\star$, aggregate TFP equals:

$$\mathcal{A} = C \cdot \mathcal{N}^{-1} = \frac{\left(\int_0^1 \left(\delta_j^S \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj},$$

where the second equality follows from the definition of δ_j^S in Equation (45). Thus:

$$\begin{aligned}\mathcal{A} &= \frac{\left(\int_0^1 \mathbb{E}\left[{\delta_j^S}^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj} \\ &= \frac{\left(\int_0^1 \mathbb{E}\left[{\delta_j^S}^{\frac{\theta-1}{\theta}} | \tau_j^\omega, \tau_j^S\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj},\end{aligned} \tag{A39}$$

where the first equality follows from the observation that δ_j^S is independent of A_j, conditional on firm j's information set $(\mu_j, \boldsymbol{\tau}_j, \boldsymbol{s}_j)$, and the second equality follows from the fact that

optimal product choice $x_j = s_j^\omega$ implies $\mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\tau_j^\omega, \tau_j^\varsigma\right]$. Using the definition of the firm-level wedge Ψ_j in Equation (45) and the optimal labor choice in Proposition 4, we have that:

$$\begin{aligned}
\mathcal{A} &= \frac{\left(\int_0^1 \Psi_j^{\frac{\theta-1}{\theta}} \cdot \mathbb{E}\left[\left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 \mathbb{E}\left[\left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj} \\
&= \exp^{\frac{\theta-1}{2}\cdot\frac{1}{\tau_\mu}} \cdot \frac{\left[\Psi\left(\underline{\tau}^\varsigma\right)^{\frac{\theta-1}{\theta}} \cdot g^R\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \left(1-\xi\left(\mu^\star\right)\right) + \Psi\left(\bar{\tau}^\varsigma\right)^{\frac{\theta-1}{\theta}} \cdot g^R\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \xi\left(\mu^\star\right)\right]^{\frac{\theta}{\theta-1}}}{g^R\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \left(1-\xi\left(\mu^\star\right)\right) + g^R\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \xi\left(\mu^\star\right)} \\
&= \overline{\Psi}\left(\mu^\star\right) \cdot \mathcal{A}(\mu^\star, g^R),
\end{aligned} \tag{A40}$$

where the two equalities follow from the definition of the information shifter $g^R(\cdot)$ in Equation (46), the expressions for $\xi(\mu^\star)$ and $\mathcal{A}(\mu^\star, g)$ in Lemma 1, as well as the definition of the macro-level wedge, $\overline{\Psi}$, in Equation (48). Given the expression for aggregate TFP, the expression for aggregate consumption then follows immediately. □

Proof of Corollary 4. See text. □

Proof of Proposition 6. Consider the problem of the social planner as described in Section 5.5. To solve this problem, we will begin with the conjecture that, given a quantity $A_j \cdot n_j$ produced by firm j, the planner can allocate consumption across the two types of consumers freely, subject to the feasibility constraint:

$$\gamma \cdot c_{hj} + (1-\gamma) \cdot c_{\ell j} \leq A_j \cdot n_j, \tag{A41}$$

where $c_{\varsigma j}$ is the consumption of variety j by type-ς household and where for now we ignore the fact that the planner also faces incentive compatibility and participation constraints. We will later show that the resulting allocations can be implemented as an outcome of a mechanism that satisfies these constraints, just as described trading mechanisms in Section 5.1.

Given this conjecture, the planner will equalize marginal utilities across types:

$$\varsigma_h^{\frac{\theta-1}{\theta}} \cdot c_{hj}^{-\frac{1}{\theta}} = \varsigma_\ell^{\frac{\theta-1}{\theta}} \cdot c_{\ell j}^{-\frac{1}{\theta}} \implies c_{\varsigma j} = \widetilde{\alpha}_j(\varsigma) \cdot Q_j, \tag{A42}$$

where $\widetilde{\alpha}_j(\varsigma)$ is the share of the good allocated to type-ς consumer, and it is the same as the share $\alpha_j\left(\varsigma|s\right)$ in Equations (39) and (40), except that the distortion term $\psi_j(s)$ is set to one.

Let λ denote the marginal value of a unit of labor to the social planner, i.e., the multiplier

on the aggregate resource constraint $\int_0^1 (n_j + \chi \cdot \iota_j) \cdot dj \leq N$. Given the above consumption allocations, it follows that the ex-post surplus (in terms of utils) produced by firm j that chooses (x_j, n_j, ι_j) can be expressed as:

$$u_j(x_j, n_j, \iota_j, \upsilon_j, \omega_j) = \frac{\theta}{\theta - 1} \cdot (\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - \lambda \cdot (n_j + \chi \cdot \iota_j) \tag{A43}$$

where δ_j is the demand shifter given by Equation (10). Thus, at *Stage 2*, the social planner's choices (x_j, n_j, ι_j) are similar to those in Proposition 1:

$$x_j = s_j^{\omega} \text{ and } n_j = \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot \widetilde{\Omega}, \tag{A44}$$

except that the effective "market size" faced by the planner is now $\widetilde{\Omega} \equiv \frac{C}{\lambda^{-\theta}}$, and:

$$\iota_j = \begin{cases} 1 & \text{if } \frac{1}{\theta-1} \cdot (\mathbb{E}[n_j|\mu_j, \bar{\boldsymbol{\tau}}] - \mathbb{E}[n_j|\mu_j, \underline{\boldsymbol{\tau}}]) \geq \chi \\ 0 & \text{otherwise} \end{cases}, \tag{A45}$$

with expected employment of firm j given by:

$$\mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j] = \exp^{(\theta-1)\cdot\mu_j} \cdot g(\boldsymbol{\tau}_j)^{\theta-1} \cdot \widetilde{\Omega}. \tag{A46}$$

Thus, at the planner's allocation, firm j produces information if and only if

$$\mu_j \geq \bar{\mu}\left(\widetilde{\Omega}\right) = \frac{1}{\theta - 1} \cdot \log\left[\frac{(\theta - 1) \cdot \chi}{\left(g(\bar{\boldsymbol{\tau}})^{\theta-1} - g(\underline{\boldsymbol{\tau}})^{\theta-1}\right) \cdot \widetilde{\Omega}}\right], \tag{A47}$$

which note is the same schedule as that given in Equation (27). Since the planner does not leave any labor unused, if we take as given the marginal type $\bar{\mu}$ that produces information, the aggregate resource constraint yields the expression for the "market size" faced by the planner:

$$\widetilde{\Omega} = \widetilde{\Omega}(\bar{\mu}) = \frac{\mathcal{A}(\bar{\mu}, g) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_{\mu}}\right)\right]}{\mathcal{A}(\bar{\mu}, g)^{\theta}}, \tag{A48}$$

where $\mathcal{A}(\cdot)$ is as in Lemma 1. This schedule is also the same as the one in Equation (32). As a result, the allocations of the social planner coincide with those of our baseline economy: the marginal type that produces information is given by the intersection of the above two schedules:

$$\mu^{\star} = \bar{\mu}\left(\widetilde{\Omega}(\mu^{\star})\right), \tag{A49}$$

while the aggregate TFP and welfare are in turn given by:

$$\mathcal{A}^\star = \mathcal{A}\left(\mu^\star, g\right) \text{ and } C^\star = \mathcal{A}^\star \cdot \left[N - \chi \cdot \Phi\left(-\mu^\star \cdot \sqrt{\tau}_\mu\right)\right]. \tag{A50}$$

As there are increasing returns to information production, we must nevertheless still verify that $\mu^\star = \arg\max_{\bar{\mu}} \mathcal{A}\left(\bar{\mu}, g\right) \cdot [N - \chi \cdot \Phi\left(-\bar{\mu}\right)]$, i.e., that the amount of information production chosen above actually maximizes welfare. But, this follows from the observation that:

$$\begin{aligned} &\frac{d}{d\bar{\mu}}\left(\mathcal{A}\left(\bar{\mu}, g\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau}_\mu\right)\right]\right) \\ &\propto (\theta - 1) \cdot \chi \cdot \frac{1}{g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1}} \cdot \frac{\mathcal{A}\left(\bar{\mu}, g\right)^{\theta-1}}{N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau}_\mu\right)} - \exp^{(\theta-1)\cdot\bar{\mu}}, \end{aligned} \tag{A51}$$

where the right-hand side is positive (negative) whenever $\bar{\mu} < \mu^\star$ ($\bar{\mu} > \mu^\star$).

Lastly, we verify the conjecture that the allocations of the planner are incentive compatible and individually rational. To this end, consider the following menu provided to consumers of each variety j by the social planner: $\mathcal{M} = \{(t_j, q_j)\}_{q_j \geq 0}$ where q_j is the quantity of the good transferred to the consumer and t_j is the transfer of the numeraire good C from the consumer to the planner—which is then rebated lump sum to the consumer. The transfer schedule in turn satisfies:

$$t_j = t\left(q_j\right) = \left[\varsigma_h^{\frac{\theta-1}{\theta}} \cdot \sigma_j^{\frac{\theta-1}{\theta}} \cdot q_j^\star(h)^{-\frac{1}{\theta}} \cdot C^{\star\frac{1}{\theta}}\right] \cdot q_j, \tag{A52}$$

where the superscript $\star$ indicates the allocations of the social planner that we obtained above (e.g., $q_j^\star(h) = \widetilde{\alpha}_j\left(h\right) \cdot A_j \cdot n_j^\star$). Note that, since the marginal utilities across consumer types are equalized, it must also be that $t_j = \left[\varsigma_\ell^{\frac{\theta-1}{\theta}} \cdot \sigma_j^{\frac{\theta-1}{\theta}} \cdot q_j^\star(\ell)^{-\frac{1}{\theta}} \cdot C^{\star\frac{1}{\theta}}\right] \cdot q_j$. Given this menu, and the conjectured equilibrium allocations, a consumer of type-ς indeed optimally chooses $q_j = q_j^\star(\varsigma)$ since it equalizes her marginal utility to her marginal cost. Therefore, by revealed preference, the planner's allocations are incentive compatible and individually rational. Note that the planner optimally ignores the information about consumer-specific idiosyncratic tastes.

We have thus shown that the social planner's allocations coincide with those of our baseline economy. Since the equilibrium of the rent-extracting economy generically differs from the equilibrium of our baseline economy, we conclude that it must generically be inefficient. □

Proof of Proposition 7. Let us begin by supposing that the social planner only has the tax instrument T at her disposal. Following in the steps of Proposition 6, we know that the marginal type $\mu^\star$ that is just indifferent to producing information in the laissez-faire

equilibrium of the rent-extracting economy is the unique maximizer of:

$$\mathcal{A}(\bar{\mu}, g^R) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right]. \tag{A53}$$

It therefore follows that the optimal information tax, which selects $\bar{\mu}$ to maximize:

$$\overline{\Psi}(\bar{\mu}) \cdot \mathcal{A}\left(\bar{\mu}, g^R\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right] \tag{A54}$$

is positive if $\overline{\Psi}'(\cdot) > 0$, i.e., segmentation is severe; and negative if $\overline{\Psi}'(\cdot) < 0$, i.e., if segmentation is mild (Definition 4). We now show that the optimal information tax is as above, provided that segmentation is socially valuable. Otherwise, if segmentation is socially destructive, the planner uses the garbling policy and the information tax is zero.

For generality, we also allow the planner to garble signals about the variety-specific productivity and demand states υ_j and ω_j, and we denote these policy instruments by $z^\upsilon \in [0,1]$ and $z^\omega \in [0,1]$ respectively. We will show that the planner does not want to use these instruments, i.e., she optimally sets $z^\upsilon = z^\omega = 0$.

Consider the parametric case where segmentation is socially destructive. Given any $\bar{\mu}$ and any garbling policy z^υ and z^ω, the product $\overline{\Psi}(\bar{\mu}) \cdot \mathcal{A}\left(\bar{\mu}, g^R\right)$ is increasing in z^ς. At the maximum when $z^\varsigma = 1$, $\overline{\Psi}(\bar{\mu})$ becomes independent of $\bar{\mu}$ and recall $\overline{\Psi}(\cdot)$ also does not depend on z^υ and z^ω. Thus, $z^\varsigma = 1$ is optimal for the planner and, given $z^\varsigma = 1$, an optimal garbling policy also sets $z^\upsilon = z^\omega = 0$, since $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\upsilon \in [0,1], z^\omega \in [0,1], z^\varsigma = 1}$ is decreasing in both z^υ and z^ω for a given $\bar{\mu}$.[42] Given the optimal garbling policy, $z^\upsilon = z^\omega = 0$ and $z^\varsigma = 1$, the optimal information tax T must be zero, since the equilibrium selects the marginal type that is the maximizer of $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\upsilon = z^\omega = 0, z^\varsigma = 1} \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right]$, which maximizes welfare, since under the optimal garbling policy, $\overline{\Psi}(\cdot)$ is independent of $\bar{\mu}$.

Consider next the parametric case where segmentation is socially valuable. In this parameter region, for any given $\bar{\mu}$, the product $\overline{\Psi}(\bar{\mu}) \cdot \mathcal{A}\left(\bar{\mu}, g^R\right)$—is now decreasing in $z^\upsilon, z^\omega, z^\varsigma$. Therefore, the social planner does not garble signals, i.e., $z^\upsilon = z^\omega = z^\varsigma = 0$, and the optimal information tax is generally non-zero (see above). □

[42] The conditioning on $(z^\upsilon, z^\omega, z^\varsigma)$ in $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\upsilon, z^\omega, z^\varsigma}$ simply indicates that we are evaluating TFP with the information shifter, in which the precision of produced information is $(\bar{\tau}^\upsilon + z^\upsilon \cdot (\underline{\tau}^\upsilon - \bar{\tau}^\upsilon), \bar{\tau}^\omega + z^\omega \cdot (\underline{\tau}^\omega - \bar{\tau}^\omega), \bar{\tau}^\varsigma + z^\varsigma \cdot (\underline{\tau}^\varsigma - \bar{\tau}^\varsigma))$.

D Model Validation and Quantification

D.1 Sectoral Misallocation

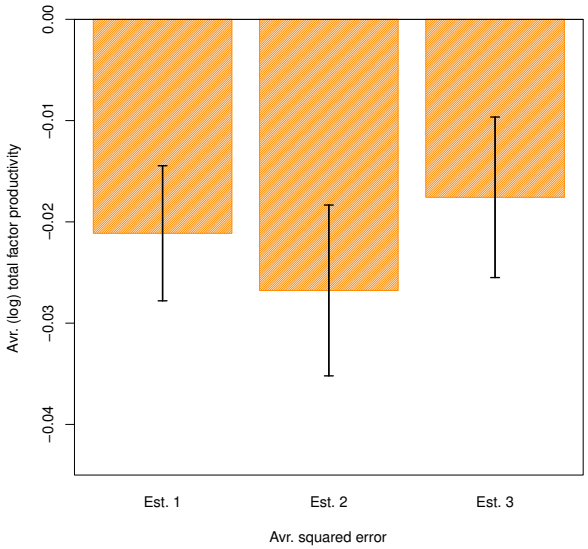
We define sectors by their four-digit NAICS industry classification. Building on the framework developed by Hsieh and Klenow (2009) and Gopinath *et al.* (2017), and consistent with our model setup, we compute our measures assuming a Cobb-Douglas production technology and monopolistic competition with CES demand. The profit-maximizing choice of an input for firm $i = \{1, 2, ..., N_s\}$ in sector $s = \{1, 2, ..., S\}$ at time $t = \{1, 2, ...\}$, thus, equates its marginal revenue product with its (sector-specific) cost. We assume the presence of two factors of productions, capital and labor. As a baseline and for comparability with our model below, we set the labor share α equal to 2/3, corresponding to the average labor share in the U.S. Our measure of misallocation is, nevertheless, not affected by the assumption that α is common across sectors, as these measures exploit within-sector variation of firm-level outcomes. Following the above terminology, we define *revenue-based total factor productivity* (TFPR) as revenue divided by output net of firm total factor productivity (TFP).[43] The *marginal revenue product of labor and capital* (MRPN and MPRK, respectively) are, by contrast, defined as revenue divided by labor and capital employed by the firm, respectively. We take revenue, labor, and capital stock measures from the I/B/E/S-Compustat sample (Appendix A.1). Panel (b) in Figure D.1 reports the average cross-sectional dispersion in mrpn, mrpk, and tfpr.[44] We report these estimates separately for accurate and inaccurate firms. We define "an informed" (good) firm as one that (i) is below the median in terms of the mean-squared-error of one-year-ahead log-revenue expectations; and (ii) one for which we have at least three observations.

[43] The production technology used by firm j is $y_{j,s,t} = A_{j,s,t} \cdot n_{j,s,t}^{\alpha} \cdot k_{j,s,t}^{1-\alpha}$, $\alpha \in (0, 1)$, where $y_{j,s,t}$ is firm output, $n_{j,s,t}$ and $k_{j,s,t}$ the amount of labor and capital employed, respectively, and $A_{j,s,t}$ is firm total factor productivity. Let $p_{j,s,t}$ be the firm-specific product price. Then *revenue-based total factor productivity* is defined as: $\text{TFP}R_{j,s,t} \equiv \frac{p_{j,s,t} \cdot y_{j,s,t}}{n_{j,s,t}^{\alpha} \cdot k_{j,s,t}^{1-\alpha}}$. The marginal revenue product of labor and capital are, by contrast: $\text{MRPN}_{j,s,t} \equiv \kappa_L \cdot \frac{p_{j,s,t} \cdot y_{j,s,t}}{n_{j,s,t}}$ and $\text{MRPK}_{j,s,t} \equiv \kappa_K \cdot \frac{p_{j,s,t} \cdot y_{j,s,t}}{k_{j,s,t}}$, where $\kappa_L \in \mathbb{R}_+$ and $\kappa_K \in \mathbb{R}_+$ are common constants that depend on the elasticity of substitution and the labor share and capital share, respectively.

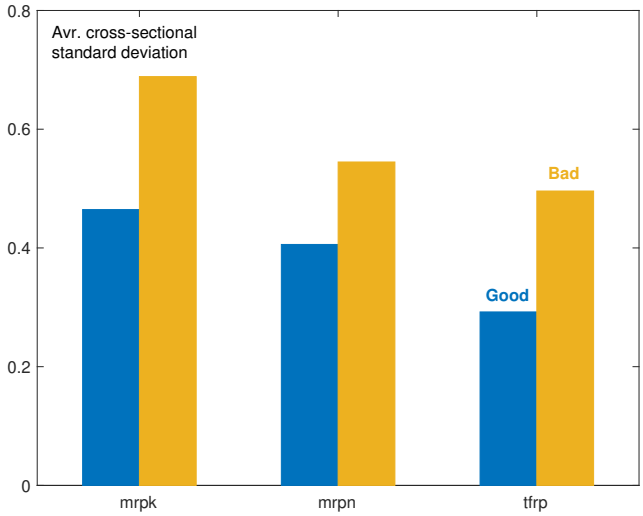
[44] We compute cross-sectional dispersion measures in two steps. First, we compute the standard deviation across firms j in a given sector s and year t. Second, for each year, we measure dispersion for the whole economy as the weighted average of dispersions across sectors. We give each sector a time-invariant weight equal to its average share in overall employment.

D.2 Qualitative Validation

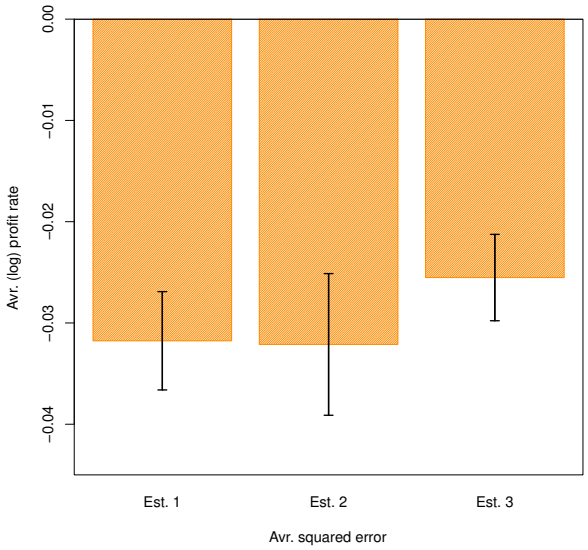
Figure D.1: Outcomes and accuracy



Panel (a): tfp-accuracy relationship



Panel (b): misallocation

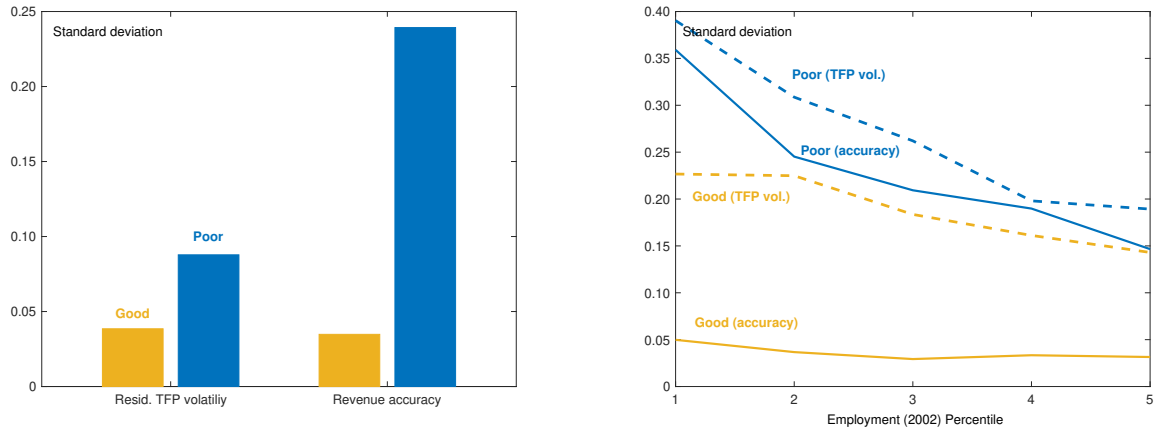


Panel (c): profit-accuracy relationship

Note: Panel (a) plots the coefficient from a regression of average firm (log-) total factor productivity on the mean-squared-error of one-year-ahead (log-)revenue forecasts (Est. 1) for the I/B/E/S-Compustat sample. The second column controls for firm age and sector fixed effects (Est. 2), while the third column trims TFP outlier observations at the 1 percent level (Est. 3). Panel (b) shows the average cross-sectional dispersion (standard deviation) in mrpn $\equiv \log(\text{MRPN})$, mrpk, and tfpr. Sectors are defined by their four-digit NAICS industry classification, and are weighted by their average share in overall employment. We define a "Good" firm as one that (i) is below the median in terms of the mean-squared-error of log-revenue expectations; and (ii) one for which we have at least three observations. We compute $\text{tfpr}_j = 1/3 \cdot \text{mrpk}_j + 2/3 \cdot \text{mrpn}_j$. Panel (c) plots the results from analogous estimates to those in Panel (a) of the average (log-)profit rate for an individual firm on its mean-squared-error and controls. Whisker intervals correspond to one-standard deviation robust (clustered) standard errors. Sample: 2002-2022.

D.3 Auxiliary Model Validation

Figure D.2: Forecast accuracy and productivity volatility



Panel (a): accuracy relationship

Panel (b): size quintiles

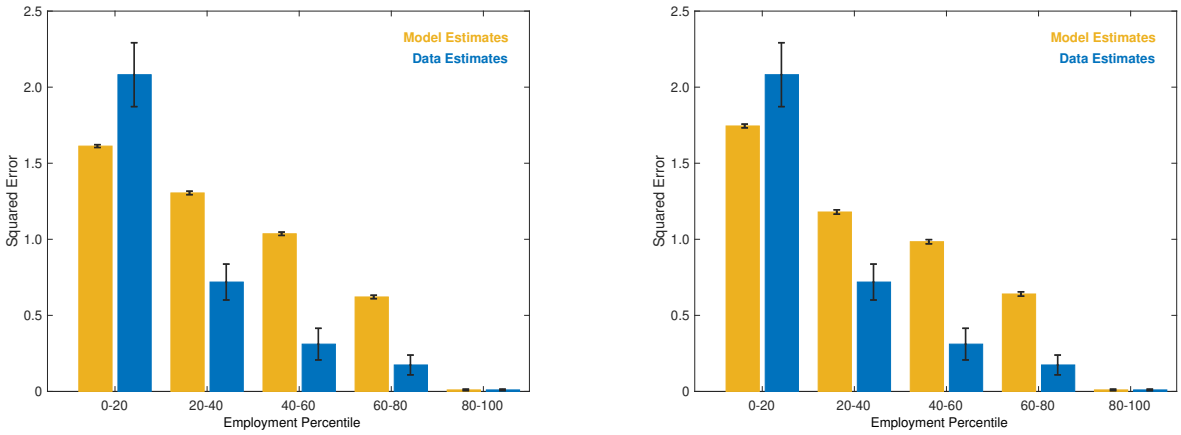
Notes: Panel (a) shows the difference between a "Good" and "Poor(ly)" informed firm's residualized (log-)tfp volatility, controlling for size, time, and and sector (NAICS-4) fixed effects. We define a "Good" firm as one that is below the 20th percentile of the mean-squared error distribution of one-year-ahead (log-) revenue forecasts. A "Poor" firm is one that is above the 80th percentile. Panel (b) shows the difference within each size quintile. Sample: 2002–2022 I/B/E/S-Compustat merger, described in Appendix A.1.

D.4 Quantitative Validation

Figure D.3: Accuracy across the size distribution

(a) Baseline economy

(b) Rent-extraction economy



Note: The figure shows the estimated relationship between squared normalized log. errors and firm size (quintiles of the initial employment distribution). We estimate this relationship both in the data and in the model (Section 2) for both the baseline economy and the rent-extracting economy. The bars labeled data control for sector and time fixed effects (Column 3 in Table I). Whisker intervals correspond to one-standard deviation robust (clustered) standard errors. Sample: 2002-2022.

D.5 Model Parametrization

Table D.1: Model parameters

Parameters	Value
Externally calibrated parameters:	
Elasticity of substitution (θ)	3.00
Labor endowment (N)	1.00
Internally calibrated parameters:	
Variance of ex-ante log-productivity (τ_μ^{-1})	1/100
Variance of log-productivity shock (τ_a^{-1})	1/7.5
Aggregate household demand shifter ($\widehat{\sigma}$)	0.03
Individual household demand shifter ($\widehat{\varsigma}$)	0.15
Low accuracy of productivity information ($\underline{\tau}^\nu$)	15.0
High Accuracy of productivity information ($\bar{\tau}^\nu$)	617.0
Low accuracy of aggregate demand information ($\underline{\tau}^\omega$)	0.50
High accuracy of aggregate demand information ($\bar{\tau}^\omega$)	0.93
Low accuracy of individual demand information ($\underline{\tau}^\varsigma$)	0.50
High accuracy of individual demand information ($\bar{\tau}^\varsigma$)	0.93
Share of $h-$types in population ($\bar{\gamma}$)	0.63
Information cost parameter (χ)	0.07
Share of information-producing firms ($\Phi(-\sqrt{\tau_\mu}\cdot\mu^\star)$)	0.10

D.6 The Extended Model

We now augment our model framework with capital and variety accumulation. To do so, we make four changes to the model. First, we assume household preferences are defined over consumption streams over time, $\{C_t^i\}$, and can be characterized by the utility function:

$$\mathcal{U}^i = \sum_{t=0}^{\infty} \beta^t \cdot \left[\log\left(C_t^i\right)\right], \tag{A55}$$

where $\beta \in (0,1)$ and C_t^i is the consumption aggregator at time t, defined consistent with Equation (1). We note that the set of goods is now endogenous in the analog of (1) and equal to $[0, M]$ for some $M > 0$, rather than fixed at $[0, 1]$. Second, we assume that each variety j at time t is produced as follows. The monopolistic firm j chooses the type x_{jt} of its variety as well as its inputs to produce quantity y_{jt} of the variety according to technology:

$$y_{jt} = A_{jt} \cdot k_{jt}^{\alpha} \cdot n_{jt}^{1-\alpha}, \tag{A56}$$

where $\alpha \in (0,1)$, and k_{jt} and n_{jt} denote the use of capital and labor inputs by firm j at time t, respectively. Capital depreciates at rate $\rho \in (0,1)$ and can be rented from households in a competitive rental market at rate r_t. Third, in this two-factor economy, we assume that information production requires the allocation of χ units of the numeraire good. Finally, we assume that the creation of a variety requires the allocation of f units of the numeraire good, and that each variety becomes obsolete at rate $\eta \in (0,1)$ in every period.

The rest of the setup is the same as before: information production yields a firm signals about productivity and demand, which are now further assumed to be iid over time. We also assume that consumer preference shocks are iid over time. Investments into capital, varieties, and information take one-period to materialize. Thus, in contrast to the framework in the main text, information production is now a form of (intangible) investment.

D.6.1 Steady State Equilibrium

We focus on the steady-state equilibrium of our framework.

Firm-level Choices. At *Stage 3*, the solution to a firm's problem is the same as before. When making its choices (x_j, n_j, k_j, ι_j), a firm faces either the baseline demand shifter δ_j or the revenue-based demand shifter $\tilde{\delta}_j^R$, depending on whether we are considering the baseline or the rent-extracting economy.[45] At *Stage 2*, the firms sets $x_j = s_j^\omega$, and chooses inputs:

$$n_j = \frac{1-\alpha}{w} \cdot \frac{w^{1-\alpha} \cdot r^\alpha}{(1-\alpha)^{1-\alpha} \cdot \alpha^\alpha} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot \tilde{g} \left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega, \tag{A57}$$

and

$$k_j = \frac{\alpha}{r} \cdot \frac{w^{1-\alpha} \cdot r^\alpha}{(1-\alpha)^{1-\alpha} \cdot \alpha^\alpha} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot \tilde{g} \left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega, \tag{A58}$$

where $\tilde{g} \in \{g, g^R\}$—depending whether we are considering the baseline or the rent-extracting economy—and where the market size faced by firms is now defined as:

$$\Omega \equiv C \cdot \left(\frac{\theta}{\theta-1} \cdot \frac{w^{1-\alpha} \cdot r^\alpha}{(1-\alpha)^{1-\alpha} \cdot \alpha^\alpha} \right)^{-\theta}. \tag{A59}$$

At *Stage 1*, firm j's expected profits for a given information choice can be expressed as:

$$\mathbb{E}\left[\pi_j | \mu_j, \boldsymbol{\tau}_j\right] = \frac{1}{\theta-1} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot \tilde{g} \left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega \cdot \frac{w^{1-\alpha} \cdot r^\alpha}{(1-\alpha)^{1-\alpha} \cdot \alpha^\alpha}, \tag{A60}$$

[45] Because of the normalization of the price index, in the extended economy, $\tilde{\delta}_j^R = \overline{\Psi}(\mu^\star) \cdot \delta_j^R$.

implying that a firm produces information if and only if

$$\beta \cdot \frac{1}{\theta - 1} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot \left(\tilde{g}\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - \tilde{g}\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \right) \cdot \Omega \geq \frac{\chi}{\frac{w^{1-\alpha}\cdot r^{\alpha}}{(1-\alpha)^{1-\alpha}\cdot\alpha^{\alpha}}}. \tag{A61}$$

Aggregate Implications. The expression for aggregate TFP now becomes:

$$\mathcal{A}^{\star} = \widetilde{\Psi}\left(\mu^{\star}\right) \cdot \mathcal{A}\left(\mu^{\star}, \tilde{g}, M\right) \tag{A62}$$

where $\widetilde{\Psi} \in \{1, \overline{\Psi}\}$—depending whether we are considering the baseline or the rent-extracting economy, with $\overline{\Psi}$ defined by Equation (48),—and:

$$\mathcal{A}\left(\mu^{\star}, \tilde{g}, M\right) = M^{\frac{1}{\theta-1}} \cdot \exp^{\frac{\theta-1}{2}\cdot\frac{1}{\tau_{\mu}}} \cdot \left(\xi\left(\mu^{\star}\right) \cdot \tilde{g}\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} + \left(1 - \xi\left(\mu^{\star}\right)\right) \cdot \tilde{g}\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \right)^{\frac{1}{\theta-1}}, \tag{A63}$$

where $\xi\left(\cdot\right)$ is given by Lemma 2. Thus, due to a "love-of-variety effect", the mass of varieties, M, also enters into the economy's TFP. To ensure that consumption growth is zero, the rental rate of capital must satisfy the condition:

$$r = \beta^{-1} - 1 + \rho. \tag{A64}$$

The wage rate must in turn be such that the labor market clears:

$$N = \mathcal{A}\left(\mu^{\star}, \tilde{g}, M\right)^{\theta-1} \cdot \Omega \cdot \left(\frac{r}{w} \cdot \frac{1-\alpha}{\alpha} \right)^{\alpha}. \tag{A65}$$

Aggregation of individual capital holdings implies that the aggregate capital stock K equals:

$$K = \mathcal{A}\left(\mu^{\star}, \tilde{g}, M\right)^{\theta-1} \cdot \Omega \cdot \left(\frac{r}{w} \cdot \frac{1-\alpha}{\alpha} \right)^{\alpha-1}, \tag{A66}$$

while aggregate consumption is:

$$C = \mathcal{A}^{\star} \cdot K^{\alpha} \cdot N^{1-\alpha} - \rho \cdot K - M \cdot \left[\chi \cdot \Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}} \right) + \eta \cdot f \right]. \tag{A67}$$

We note that the expression for C includes both the costs of information production and of capital and variety creation. Finally, in equilibrium, as all potential entrant varieties are ex-ante identical, it follows that expected profits (inclusive of entry costs) must equal to zero:

$$\chi \cdot \Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}} \right) + (1 - \beta \cdot (1 - \eta)) \cdot f = \frac{1}{\theta - 1} \cdot \frac{\mathcal{A}\left(\mu^{\star}, \tilde{g}, M\right)^{\theta-1}}{M} \cdot \beta \cdot \Omega \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}}. \tag{A68}$$

Equations (A57) through (A68) fully characterize the steady state of the extended economy.

D.6.2 Extended Calibration

We set the fixed cost of entry, f, so that the mass of firms in the steady-state of the model equals one (i.e., the mass in the rent-extracting and baseline economy). We set the discount factor (β), the capital depreciation rate (ρ), the capital share (α), and the exit rate (η) to standard values used in the literature (0.96, 0.02, 0.33, 0.02, respectively). The rest of the parameters are calibrated in the same manner as in the rent-extracting and baseline economy.

D.6.3 Extension and Decomposition

Table D.2: Extended model: decomposition of the rise in TFP

Model	Overall (%)	Scale	Product	Pricing	Variety
Baseline	5.0	2.1	2.3	0.0	0.5
Rent-extraction	2.9	1.6	1.8	-0.5	-0.1

Note: The table decomposes the rise in TFP in Figure 9 into its four constituent channels: (i) scale, (ii) product design, (iii) pricing, and (iv) "variety creation" using Equation (A62). The table does so for the best-case and for the worst-case economy identified as in the main text.

References

Abis, S. and Veldkamp, L. (2024). The changing economics of knowledge production. *The Review of Financial Studies*, **37** (1), 89–118. 3.3

Adams, J. J., Fang, M., Liu, Z. and Wang, Y. (2025). The rise of ai pricing: Trends, driving forces, and implications for firm performance. *Journal of Monetary Economics*, p. 103875. 3, 3.3, 5.1, 6.1

Ali, S. N., Lewis, G. and Vasserman, S. (2020). Voluntary disclosure and personalized pricing. *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 537–538. 3.3

Angeletos, G.-M., Huo, Z. and Sastry, K. A. (2021). Imperfect macroeconomic expectations: Evidence and theory. *NBER Macroeconomics Annual*, **35** (1), 1–86. 2, 1

—, Iovino, L. and La'o, J. (2016). Real rigidity, nominal rigidity, and the social value of information. *American Economic Review*, **106** (01), 200–227. 3.2

— and Lian, C. (2016). Incomplete information in macroeconomics: Accommodating frictions in coordination. **2**, 1065–1240. 2

— and Pavan, A. (2007). Efficient Use of Information and Social Value of Information. *Econometrica*, **75** (4), 1103–1142. 1

Arias, A., Hansen, G. D. and Ohanian, L. E. (2007). Why have business cycle fluctuations become less volatile? *Economic theory*, **32**, 43–58. 2

Asriyan, V., Fuchs, W. and Green, B. (2017). Information spillovers in asset markets with correlated values. *American Economic Review*, **107** (7), 2007–2040. 7

—, — and Lorecchio, C. (2025). Multilateral bargaining with information spillovers. In *Technical Report, Working Paper*. 1

Autor, D., Dorn, D., Katz, L. F., Patterson, C. and Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, **135** (2), 645–709. 9

Babina, T., Fedyk, A., He, A. and Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, **151**, 103745. 3

Baley, I. and Veldkamp, L. (2025). *The Data Economy: Tools and Applications*. Princeton University Press. 1, 4.3, 6.1

Begenau, J., Farboodi, M. and Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, **97**, 71–87. 1

Blanchard, O. J., L'Huillier, J.-P. and Lorenzoni, G. (2013). News, noise, and fluctuations: An empirical exploration. *American Economic Review*, **103** (7), 3045–70. 2

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I. and Van Reenen, J. (2019). What drives differences in management practices? *American Economic Review*, **109** (5). 2

—, Floetotto, M., Jaimovich, N., Saporta-Eksten, I. and Terry, S. J. (2018). Really uncertain business cycles. *Econometrica*, **86** (3), 1031–1065. 2

Bornstein, G. and Peter, A. (2024). Nonlinear pricing and misallocation. 6, 21, 6.1, II, C.1

Brynjolfsson, E., McAfee, A., Sorell, M. and Zhu, F. (2008). Scale without mass: business process replication and industry dynamics. *Harvard Business School Technology & Operations Mgt. Unit Research Paper*, (07-016). 2

— and McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, **106** (5), 133–139. 1, 3.3, 6.1

— and — (2024). The rapid adoption of data-driven decision-making. *Working Paper*. 1, 6.1

Chahrour, R. and Jurado, K. (2018). News or noise? the missing link. *American Economic Review*, **108** (7), 1702–36. 2

Chen, C., Hattori, T. and Luo, Y. (2023). Information rigidity and elastic attention: Evidence from japan. 4, 8, A.4

Chen, H., Li, X., Pei, G. and Xin, Q. (2024). Heterogeneous overreaction in expectation formation: Evidence and theory. *Journal of Economic Theory*, p. 105839. 4, 8, A.4

Chiavari, A. and Goraya, S. (2023). The rise of intangible capital and the macroeconomic implicationsn. *mimeo*. 2, A.1, A.12

Coyle, D. and Hampton, L. (2024). 21st century progress in computing. *Telecommunications Policy*, **48** (1), 102649. 5, 4.3

Crémer, J. and McLean, R. P. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pp. 1247–1257. 22

David, J. M., Hopenhayn, H. A. and Venkateswaran, V. (2016). Information, misallocation, and aggregate productivity. *The Quarterly Journal of Economics*, **131** (2), 943–1005. 1, 4.1

— and Venkateswaran, V. (2019). The sources of capital misallocation. *American Economic Review*, **109** (7), 2531–2567. 1, 4.1, 6.1

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American economic review*, **67** (3), 297–308. 1, 26, 5.5

Eeckhout, J. and Veldkamp, L. (2025). Data and markups: a macro-finance perspective. *UPF Working Paper*. 1, 19

European Commission, E. (2020). A european strategy for data. *Strategy report*. 5

Farboodi, M., Haghpanah, N. and Shourideh, A. (2025). Good data and bad data: The welfare effects of price discrimination. *arXiv preprint arXiv:2502.03641*. 1

— and Veldkamp, L. (2020). Long-run growth of financial data technology. *American Economic Review*, **110** (8), 2485–2523. 1

— and — (2024). *A model of the data economy*. Tech. rep., National Bureau of Economic Research Cambridge, MA, USA. 1, 4.1

Feng, Y., Zhao, Y., Zheng, H., Li, Z. and Tan, J. (2020). Data-driven product design toward intelligent manufacturing: A review. *International Journal of Advanced Robotic Systems*, **17** (2). 3

Fudenberg, D. and Villas-Boas, J. M. (2012). Price discrimination in the digital economy. 1, 5.1

Gill, S. S., Wu, H., Patros, P., Ottaviani, C., Arora, P., Pujol, V. C., Haunschild, D., Parlikad, A. K., Cetinkaya, O., Lutfiyya, H. *et al.* (2024). Modern computing: Vision and challenges. *Telematics and Informatics Reports*, p. 100116. 5, 4.3

Goldfarb, A., Greenstein, S. M. and Tucker, C. E. (2015). *Economic analysis of the digital economy*. University of Chicago Press. 1

Gopinath, G., Kalemli-Özcan, Ş., Karabarbounis, L. and Villegas-Sanchez, C. (2017). Capital allocation and productivity in south europe. *The Quarterly Journal of Economics*, **132** (4), 1915–1967. D.1

Gorton, G. and Ordonez, G. (2014). Collateral crises. *American Economic Review*, **104** (2), 343–378. 7

Graham, J., Meyer, B., Parker, N. and Waddel, S. (2023). The cfo survey. *Duke University mimeo*. 2

Hannak, A., Soeller, G., Lazer, D., Mislove, A. and Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*, pp. 305–318. 5.1

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, **124** (4), 1403–1448. 6.1, D.1

IDC, I. D. C. (2021). Idc data wallet tracker, 2021. *IDC Working Paper*. 1, 6.1

Jin, W. and McElheran, K. (2024). Economies before scale: It strategy and performance dynamics of young us businesses. *Management Science*. 3

Kehoe, P. J., Larsen, B. J. and Pastorino, E. (2018). Dynamic competition in the era of big data. In *Technical Report, Working Paper*, Stanford University Stanford, CA. 1

Kwon, S. Y., Ma, Y. and Zimmermann, K. (2023). 100 years of rising corporate concentration. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-20). 9

Lorenzini, L. and Martner, A. (2026). Aggregate outcomes of nonlinear prices in supply chains. 6, 21

Lorenzoni, G. (2009). A theory of demand shocks. *American economic review*, **99** (5), 2050–2084. 1

Lucas, R. E. (1977). Understanding business cycles. *Essential readings in economics*, pp. 306–327. 2

Lucas, R. E. J. (1972). Expectations and the neutrality of money. *Journal of Economic Theory*, **4** (2), 103–124. 1, 3.3

Maćkowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *The American Economic Review*, **99** (3), 769–803. 1

Malherbe, F. (2012). Market discipline and securitization. 7

Mankiw, N. G. and Reis, R. (2002). Sticky information versus sticky prices: a proposal to replace the nesw keynesian phillips curve. *The Quarterly Journal of Economics*, **117** (4), 1295–1328. 1

Mckinsey and Company (2023). Frontiers of the data economy? 1

Nordhaus, W. D. (2008). Two centuries of productivity growth in computing. *The Journal of Economic History*, **67** (1), 128–159. 5, 4.3

OECD (2018). MS Windows NT kernel description. 5, 5.1

O'Neill, L. (2023). 10 companies that are using big data. *ICAS Working Paper*. 3, 13, 19

Ordonez, G. (2009). *Larger crises, slower recoveries: the asymmetric effects of financial frictions*. Citeseer. 1

Ottonello, P. and Winberry, T. (2020). Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, **88** (6), 2473–2502. 31, II, A.1, A.11

Senga, T. (2018). A new look at uncertainty shocks: Imperfect information and misallocation. *Working paper*. 4

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50** (3), 665–690. 2

Tanaka, M., Bloom, N., David, J. M. and Koga, M. (2020). Firm performance and macro forecast accuracy. *Journal of Monetary Economics*, **114**, 26–41. 4, 8, A.4

Varian, H. R. (1989). Chapter 10 price discrimination. *Handbook of Industrial Organization*, vol. 1, Elsevier, pp. 597–654. 1

Veldkamp, L. and Chung, C. (2024). Data and the aggregate economy. *Journal of Economic Literature*, **62** (2), 458–484. 3.3

Woodford, M. (2002). *Imperfect Common Knowledge and the Effects of Monetary Policy*. Nber working papers, Department of Economics, Columbia University. 2, 1

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foster, L. and Dinlersoz, E. (2021). Advanced technologies adoption and use by us firms: Evidence from the annual business survey. *NBER Working Paper*. 3