

Where Has All the Data Gone?

Maryam Farboodi

MIT Sloan

Adrien Matray

Princeton University

Laura Veldkamp

Columbia Business School

Venky Venkateswaran

Stern School of Business, NYU

Since the finance industry is transforming into a data industry, measuring the quantity of data investors have about various assets is important. Informed by a structural model, we develop such a cross-sectional measure. We show how our measure differs from price informativeness and use it to document a new fact: data about large high-growth firms is becoming increasingly abundant, relative to data about other firms. Our structural model offers an explanation for this data divergence: large high-growth firms' data became more valuable, as big firms got bigger and growth magnified the effect of these changes in size. (JEL G10, G11, G12, G14)

Received March 25, 2020; editorial decision May 6, 2021 by Editor Itay Goldstein.

Data are becoming more central to the practice of finance. To address the myriad of questions that are arising about data value and data choice, we need a quantitative measure of data being used by market participants. Firms want to know: how much do others know about various types of assets? This paper develops a data measure to answer this question.

The challenges of measuring data processing are manifold. For one, it is not directly observable. While some of it is bought and sold, much of it is not. Proxies are available: like counts of news stories, information technology expenditures, or analyst coverage. These are suggestive, but are quite crude,

We thank John Barry, Matias Covarrubias, Ye Zhen, and Joseph Abadi for their excellent research assistance; Yifeng Guo, Vincent Glode, Brian Weller, Ben Golub, Pete Kyle, and Liyan Yang for their help and suggestions; seminar participants at Wharton, Columbia, and MIT; and participants in the 2017 NBER risk group, 2018 Econometric Society meetings, 2019 AFA, and 2020 Georgia FinTech conference for their comments. Send correspondence to Laura Veldkamp, lv2405@columbia.edu.

The Review of Financial Studies 35 (2022) 3101–3138

© The Author(s) 2021. Published by Oxford University Press on behalf of The Society for Financial Studies. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

doi: 10.1093/rfs/hhab110

Advance Access publication October 17, 2021

especially if one wants a precise answer to the question: how much information about stocks did investors extract through data processing? Measures of the information contained in market prices reflect the amount of data, and are influenced by market volatility or the price sensitivity to data, factors that differ across assets.

We address these challenges by building a simple structural model to guide our measurement. The model shows how data are related to and yet distinct from concepts like price informativeness. It also provides a formula to correct a price information measure for the effect of asset characteristics and obtain a pure measure of data.

Next, we use this toolkit to study cross-sectional patterns in data in the U.S. equity market over the past few decades. We group assets by size and growth prospects: we chose these dimensions because they drive the value of data processing in the model. Our analysis reveals a new fact: diverging trends in data processing across different assets. Investors in large high-growth firms are basing their decisions on more and more data. For other assets, data appear stagnant, in comparison. In other words, ever-growing reams of financial data may be helping price assets more accurately. But this additional data might not deliver financial efficiency benefits for the vast majority of firms. This divergence is consistent with reduced-form measures, like price informativeness measures and analyst coverage patterns, with different magnitudes. However, quantifying the magnitude of the divergence in units of data precision is valuable, beyond the reduced-form evidence.

The third contribution of the paper is to explore data valuation. We find that the value of data depends on firm size and growth. This finding is what motivated us to sort firms by size and growth in the empirical analysis. These are dimensions along which data choices should vary. When we use size and growth estimates to quantify data value, we uncover a potential explanation for data divergence: the value of large, high-growth firm data has diverged, as large firms have grown relatively larger.

Section 1 begins with a simple model designed to relate data precision to observables. Our theoretical framework is a standard noisy rational expectations framework with multiple assets. The theory points to a particular moment as a natural starting point for our analysis of data processing: estimate the coefficient for prices in a regression of future cash flows on a constant, prices and controls. This coefficient, referred to as price informativeness by papers like Bai, Philippon, and Savov (2016) measures how closely prices reflect future firm outcomes. This is obviously affected by the amount of data processed but also depends on other firm characteristics, making trends in this variable difficult to interpret or attribute solely to changes in data processing. Our model overcomes this difficulty: it offers a simple expression that relates the price informativeness measure to data, in a way that holds with minimal theoretical assumptions. Specifically, it can be decomposed into components that depend on data processing, cash flow growth, and volatility. That cash flow and volatility

can be directly estimated from financial market observables allows us to back out a precise measure of data processed by investors.

Section 2 provides detail of how we estimate our model structurally. This includes the description of our sample, our variable construction, as well as the moments used for the structural estimation.

In Section 3, we report how informativeness of prices changes across different classes of assets and decompose that change into changes in volatility, growth and data. We find that, over the last 50 years, data about most firms has stagnated. However, one category of data has become much more abundant: information about large high-growth firms. Strikingly, while firm growth and volatility have also changed over time, their changes work against this trend for the most part. For example, by themselves, they would imply falling price informativeness for large, high-growth firms as well. Thus, our measurement exercise reveals that data divergence is the key to understanding the changes over the last few decades. We also contrast our measure, both theoretically and quantitatively, with other measures, such as price informativeness, comovement, and absolute price informativeness.

Finally, Section 4 uses the model to explore the underlying drivers of this rising abundance of data processing on large, high-growth firms. Specifically, we compute a model implied value of data, which is increasing in the size of the firm, volatility of its cash flows and growth prospects. While this is not surprising *per se*, the model yields a simple formula that shows exactly how these characteristics interact, offset, and amplify each other. This measure allows us to precisely rank assets based on the value of learning about them, which then predicts the types of assets are learned about.

We find that value of data about large high-growth stocks has diverged in recent years, which offers a potential explanation for the trends we see in data processing. A key factor behind this divergence is a similar pattern in firm size: large firms got much larger, compared to small firms. Since an increase in size allows investors to take larger positions based on their data processing, the divergence in size makes larger firms even more attractive to learn about. Finally, growth amplifies changes in the value of data. Growth multiplies size in data value. Although data for all large firms became more valuable as large firms got bigger, in most decades, this effect was strongest for the large high-growth firms. The fact that our prediction about the value of the data is consistent with the data patterns we see, both offers an explanation for our facts and bolsters our confidence in our measurement approach.

Thus, as overall data processing capacity increased in the economy, most of it seems to have gone to learning about the prospects of large high-growth firms. Other types of firms benefited little from this data revolution.

Our methodology is most related to Bai, Philippon, and Savov (2016) and Davila and Parlato (2016b), who propose measures of price informativeness.

Their measure captures the ability of prices to forecast or aggregate information. Such a measure is valuable because it may relate to real efficiency.¹

Similarly, measures of comovement, synchronicity, or R^2 (Durnev, Morck, and Yeung 2004) quantify aggregate price variation, relative to stock-specific price variation. Our question differs. We want to know how the allocation of financial data precision, across asset types, has changed over time. Our measure is valuable because we need it to value or choose data. Section 2.3 compares these measures and reveals important differences: noise, size, and growth all drive a wedge between the previously used price informativeness or comovement measures and our data measure. We measure these wedges and find they are quantitatively large. Finally, previous exercises did not explain why trends emerged. Our approach does.

Empirical work in this area primarily uses proxies for data or information, such as news consumption (Ben-Rephael et al. 2021), social media text (Ranco et al. 2015), analyst coverage (Hong and Kacperczyk 2010; Kelly and Ljungqvist 2012), or earnings announcements (Martineau 2017). These papers measure the effect of a particular information channel and for the most part, are interested in cross-sectional determinants rather changes over time. Our goal is to measure all the information investors use, from all channels, and to document how that has changed over time.

Work by Stambaugh (2014) and Glode, Green, and Lowery (2012) does explain the reason for overall information trends. But their focus is on aggregate trends that affect all assets. These authors highlight forces, such as rising institutional ownership and indexation. Such forces could be incorporated into our measurement framework by changing the marginal benefit of all firms' data. But our focus is on why these trends differ across asset classes and what part of that change is information versus divergent asset characteristics.

Finally, the way in which we model data has its origins in information theory/computer science, and is similar to work on rational inattention (Sims 2003; Maćkowiak and Wiederholt 2009; Kacperczyk, Nosal, and Stevens 2019). Similar equilibrium models with information choice have been used to explain income inequality (Kacperczyk, Nosal, and Stevens 2019), information aversion (Andries and Haddad 2020), home bias (Mondria, Wu, and Zhang 2010; Van Nieuwerburgh and Veldkamp 2009), and mutual fund returns (Pástor and Stambaugh 2012), among other phenomena. Related microstructure work explores the frequency of information acquisition and trading (Kyle and Lee 2017; Dugast and Foucault 2018; Chordia, Green, and Kottimukkalur 2018; Crouzet, Dew-Becker, and Nathanson 2020). Empirical work in this

¹ An extensive literature tackles how asset price informativeness affects real investment. Ozdenoren and Yuan (2008), Bond and Eraslan (2010), Goldstein, Ozdenoren, and Yuan (2013), David, Hopenhayn, and Venkateswaran (2016), Dow, Goldstein, and Guembel (2017), Dessaint et al. complement our work by showing how the financial information trends we document could have real economic effects. Bond, Edmans, and Goldstein (2012b) review this literature and conclude that the relationship between market efficiency and real efficiency is not necessarily monotone and depends on the environment.

vein Katz, Lustig, and Nielsen (2017) finds evidence of rational inattention like information frictions in the cross-section of asset prices. What we add to this literature is using the theory for structural estimation. Our structure allows us to distinguish changes in information from changes in asset characteristics.

1. A Structural Framework for Data Measurement

The main objective of the paper is to develop a measure of investors' data precision from asset prices. This is related to measures of price informativeness. But we know that informativeness also reflects differences in price-earnings ratios, related to firm growth, and differences in firm price volatility. One approach would be to simply control for such asset characteristics in a linear regression. However, one problem with this approach is that growth and volatility themselves affect the value of collecting and processing data and therefore, are likely to be correlated with investors' data. As such, adding them as controls can remove some of what we hope to measure. Another problem is nonlinearity: the effect of growth, for example, is probably not additive. In fact, this is exactly what happens in our model, where it interacts with the measure of data multiplicatively. Furthermore, our goal is to develop a measure of data processing that guides the choice of valuation of data by investors. To do so, it needs to be consistent with—or interpretable in terms of—a valuation or data portfolio choice model. For all these reasons, we turn to a structural approach to inform us about how to properly measure data.

We work with the simplest theoretical framework that achieves this objective. The setup is a standard noisy rational expectations model with multiple assets, in the spirit of Admati (1985) and Van Nieuwerburgh and Veldkamp (2009). The model yields simple, intuitive expressions for the objects of interest, including a measure of price informativeness, as a function of both asset characteristics and investor data. These expressions form the basis for an empirical strategy that disentangles asset characteristics from investor data, using observable moments of stock prices and cash flows.

Model A unit measure of investors trade multiple stocks (indexed by f). We assume that these assets belonging to different groups (indexed by j), where assets within a group share a number of parameters. The empirical analogs and the rationale for choosing will be described in detail in Section 2.2. A share is a claim to a stream of dividends. Dividends grow at different rates across groups. We denote the group-specific growth rate by g_j . The flow dividend of stock f in group j in period 1 has two random innovations: one correlated across firms and the other idiosyncratic (i.e., stock specific). These are denoted by $\bar{\epsilon}_{fj1}$ and ϵ_{fj1} , respectively. For our baseline analysis, we will assume that the correlated innovation has a one-factor structure; that is, it is the product of a firm-specific loading and the realization of an aggregate factor: $\bar{\epsilon}_{fj1} = \tilde{\beta}_{f,j} \bar{\epsilon}_1$, where $\tilde{\beta}_{f,j}$ is

the firm-specific loading.² The idiosyncratic component is normally distributed with a zero mean. Formally,

$$d_{fj1}^* = g_j d_{fj0}^* + \bar{\epsilon}_{fj1} + \epsilon_{fj1}, \quad \epsilon_{fj1} \sim N(0, \Sigma_{jd}). \quad (1)$$

The dividends for periods $s=2, 3, \dots$ are given by $d_{fjs}^* = g_j^{s-1} d_{fj1}^*$.

The assumption of no residual uncertainty after period 1 is only for simplicity. It implies that the value of the stock at the end of period 1 is given by³

$$V_{fj1}^* \equiv \sum_{s=1}^{\infty} \frac{d_{fjs}^*}{r^s} = \frac{r}{r - g_j} d_{fj1}^*. \quad (2)$$

where r is the riskless rate. Note how g_j enters the factor that determines the earnings-to-valuation ratio. This will be helpful later on for the interpretation of g_j as growth.

Supply The supply of each asset has a (commonly known) asset-specific mean \bar{x}_{fj} as well as an unobserved random component $\tilde{x}_{fj} \sim N(0, \Sigma_{jx})$. Assets within a group have the same mean supply, that is, $\bar{x}_{fj} = \bar{x}_j$. Formally, the total supply of asset f in group j is $\bar{x}_j + \tilde{x}_{fj}$ shares. Thus, as with the cash flow process, parameters driving asset supply are group specific.

Preferences and portfolio choice. Investors, indexed by i , are endowed with an initial wealth \bar{W}^i and mean-variance preferences over their end-of-period wealth.

At the start of period 1, investors make portfolio choices, conditional on an information set \mathcal{I}^i . Formally, investor i with absolute risk aversion ρ_i chooses $\{q_j^i\}$, the number of shares of asset j , to solve:

$$\max_{\{q_{fj}^i\}} \mathbb{E}[U^i | \mathcal{I}^i] = \max_{\{q_{fj}^i\}} \rho_i \mathbb{E}[W^i | \mathcal{I}^i] - \frac{\rho_i^2}{2} \text{Var}(W^i | \mathcal{I}^i), \quad (3)$$

$$\text{where} \quad W^i = r \bar{W}^i + \sum_j \sum_f q_{dj}^i (V_{fj1}^* - r P_{fj1}^*).$$

r is the riskless rate, P_{fj1}^* is the equilibrium market clearing price of asset f in group j and V_{fj1}^* is the present discounted asset value from (2). At the end of the period, d_{fj1} is observed, investors sell their holdings and consume.

This mean-variance representation is a simple way to a broad array of preference specifications. For example, the coefficient of absolute risk aversion ρ_i is allowed to be any nonrandom function of initial wealth, \bar{W}^i . Thus,

² Appendix B.4 shows that our results hold under a more complicated, group-level multifactor structure.

³ An obvious alternative assumption is that all uncertainty is not resolved at the end of period 1 and investors sell their assets at a market price, which depends, among other things, on the information of future participants, as in Farboodi and Veldkamp (2020). This delivers a similar solution, except that the dependence on future information introduces another fixed point problem, which complicates the analysis considerably, without providing additional insight.

these preferences could be derived from decreasing absolute risk aversion preferences, or even constant relative risk aversion, in initial wealth.

Information. Our focus is on data used to pick stocks, rather than for timing the overall market. This focus is motivated by our interest in cross-asset differences, which empirically seem to be driven mostly by stock-specific factors. In our sample, more than 90% of the variation in prices is stock specific. Moreover, time variation in cross-sectional moments is easier to precisely estimate.⁴

With this goal in mind, we make the simplifying assumption that all investors know the common component of the asset payoffs (i.e., the aggregate factor $\bar{\epsilon}_1$). This assumption, along with the structure of payoffs and preferences, allows us to analyze asset-specific learning without making further assumptions on the distribution of the common component.

For each risky asset f in group j , investor i privately observes k_j^i data points. We call k_j^i investor i 's *net private data* about asset j . Each data point is a noisy private signal (with errors that are iid across assets and investors) of the end-of-period asset-specific cash flow ϵ_{fj1} :⁵

$$\eta_{fj}^{i,m} = \epsilon_{fj1} + e_{fj}^{i,m}, \quad e_{fj}^{i,m} \sim_{iid} N(0, 1),$$

for $m \in \{1, \dots, k_j^i\}$. The average amount of private data about asset j in the market is

$$K_j = \int k_j^i di. \quad (4)$$

In addition, investors also observe the realized market-clearing price P_{fj1}^* (characterized later) and also optimally incorporate the information contained in that price. Thus, investor i 's information set, for asset f in group j , consists of the dividend realization in period 0, a set of private signals, and the market-clearing price: $\mathcal{I}^i = \{\{d_{fj0}^*\}, \{\eta_{fj}^{i,m}\}_{m=1}^{k_j^i}, \{P_{fj1}^*\}\}$. We conjecture (and later verify) that the information in the market price can be expressed as a signal of the cash flow innovation, ϵ_{fj1} with additive Gaussian noise. Then, Bayes' law for normally distributed random variables yields the following expression for investor i 's precision about the cash flow d_{fj1}^* of any assets in group j , denoted by $(\Sigma_j^i)^{-1}$:

$$(\Sigma_j^i)^{-1} \equiv \text{Var}[\epsilon_{fj1} | \mathcal{I}^i]^{-1} = \Sigma_{jd}^{-1} + (\Sigma_{jp}^i)^{-1} + k_j^i, \quad (5)$$

⁴ Having said that, we believe that one could easily adapt the framework and the empirical strategy to measure data about aggregate factors instead.

⁵ This language suggests discrete numbers of signals. Since working with discrete variables complicates the analysis considerably and adds little insight, we treat k_j^i as a continuous variable. Formally, we can take a quasi-continuous limit. If each data point has variance α , this limit takes the number of data points to be αk_j^i and then sends $\alpha \rightarrow \infty$. At the limit, the precision of the set of signals becomes continuous.

where $(\Sigma_{jp}^i)^{-1}$ is the precision of the market price signal (to be characterized later). This notation allows for the possibility that different investors learn differently from market prices. This could occur, for example, if it was costly to extract information from prices. The symmetric case, with $(\Sigma_{jp}^i)^{-1} = \Sigma_{jp}^{-1}$ is a natural starting point and is maintained in our characterization of equilibrium below.

The average marketwide precision, denoted by $(\bar{\Sigma}_j)^{-1}$, is

$$\begin{aligned} (\bar{\Sigma}_j)^{-1} &= \int (\Sigma_j^i)^{-1} di = \Sigma_{jd}^{-1} + \int (\Sigma_{jp}^i)^{-1} di + \int k_j^i di \\ &= \Sigma_{jd}^{-1} + \Sigma_{jp}^{-1} + K_j. \end{aligned} \quad (6)$$

where Σ_{jp}^{-1} and K_j are (marketwide) averages of the precision gained from the price signal and net private data, respectively.

Equilibrium A rational expectations equilibrium is a set of functions for prices P_{fj1}^* , and portfolio choices q_{fj}^i such that (a) given the induced information sets \mathcal{I}^i , the portfolio choices solve (3), and (b) markets clear, that is, $\forall f, j, \int q_{fj}^i di = \bar{x}_j + \tilde{x}_{fj}$.

To solve for the equilibrium, we conjecture a linear form for the price function and solve for the corresponding coefficients. We relegate the details to the appendix and present the solution in the following result:

Proposition 1. In equilibrium, the price of asset j is given by

$$r P_{fj1}^* = A_{fj} + B_j \epsilon_{fj1} + C_j \tilde{x}_{fj}, \quad (7)$$

$$\text{where} \quad A_{fj} = \bar{P}_{fj1} + \left(\frac{r}{r - g_j} \right) g_j d_{fj0}^* - \bar{\rho} \left(\frac{r}{r - g_j} \right)^2 \bar{\Sigma}_j \tilde{x}_j, \quad (8)$$

$$B_j = \frac{r}{r - g_j} \left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right), \quad (9)$$

$$C_j = - \left(\frac{r}{r - g_j} \right)^2 \bar{\Sigma}_j \left(\frac{K_j \Sigma_{jx}}{\bar{\rho}} + 1 \right), \quad (10)$$

$$\Sigma_{jp}^{-1} = \left(\frac{B_j}{C_j} \right)^2 \Sigma_{jx}^{-1}. \quad (11)$$

$\bar{\rho}^{-1} := \bar{\Sigma}_j \int \rho_i^{-1} (\Sigma_j^i)^{-1} di$ is a precision-weighted average of investors' risk tolerance.⁶ The term \bar{P}_{fj1} captures the valuation of the common component of dividends ($\bar{\epsilon}_{fj}$).

⁶ Assuming $\bar{\rho}$ is constant across assets amounts to assuming that risk tolerance and precision are either uncorrelated or do not covary differently for different assets.

Equation (9) shows that the coefficient for current innovations to cash flows, B_j , is the usual Gordon growth factor, $\frac{r}{r-g_j}$, adjusted by a factor $\left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}}\right)$. This factor captures the effects of data processing by investors, thus we call it *data*. If investors do not have data about asset j (apart from their prior), then the average posterior variance $\bar{\Sigma}_j$ is equal to the prior variance Σ_{jd} , and the coefficient $B_j=0$. In other words, the price cannot possibly reflect information that no investor has learned. At the other extreme, if the average investor is perfectly informed about current cash flows, then $\bar{\Sigma}_j=0$ and $B_j=\frac{r}{r-g_j}$, the Gordon growth factor. Thus, the extent to which the stock price covaries with cash flow innovations is informative about how much data related to asset j is processed by the average investor.

Equation (11) characterizes the precision of the price as a signal of future dividends. The linear form of the equilibrium price implies that it is informationally equivalent to $\frac{rP_{fj1}^* - A_{fj}}{B_j} = \epsilon_{fj1} + \frac{C_j}{B_j} \tilde{x}_{fj}$, that is, a noisy signal of the innovation to cash flows with a precision $\left(\frac{B_j}{C_j}\right)^2 \Sigma_{jx}^{-1}$. The signal is more precise when the sensitivity of the equilibrium price to fundamentals relative to supply noise (B_j/C_j) is high, or the variance of supply Σ_{jx} is low.

Next, we construct a moment, which we term stock-specific price informativeness, or *PINF*, that will guide our empirical strategy in the following section. Formally, we define s -period-ahead stock-specific price informativeness of group- j as

$$PINF_{js} \equiv \frac{\text{Cov}(d_{fjs}^*, P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)}{\text{StdDev}(P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)}. \quad (12)$$

This moment captures the extent to which the stock-specific components of current prices and cash flows s periods ahead covary with each other. As we will see in the next section, this can be easily estimated with a simple linear regression using data on market capitalization, cash flows and assets.

Our framework implies that $PINF_{js}$ can be expressed as follows:

$$PINF_{js} = \underbrace{\frac{\Sigma_{jd}}{\text{StdDev}(P_{fj1})}}_{\text{volatility}} \underbrace{\frac{g_j^s}{r-g_j}}_{\text{growth}} \underbrace{\left[1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}}\right]}_{\text{data}}. \quad (13)$$

where $P_{fj1} = rP_{fj1}^* - A_{fj}$ is the component of prices that pertains to the stock-specific innovation, ϵ_{fj1} . Equation (13) forms the core of our analysis. It reveals that $PINF_{js}$ can be decomposed into three parts. We term the first component *volatility*: it is the ratio of the variability of cash flow innovations to that of prices. All else equal, an asset whose prices are more volatile (relative to cash flows) will exhibit a lower degree of informativeness.⁷

⁷ This insight also appears in Dávila and Parlato (2019).

The second component is related to *growth*. Intuitively, a faster growing cash flow process implies that prices load on current cash flows to a greater extent. This increases their covariance and contributes to a higher *PINF*. In our structure, growth prospects (or equivalently, the cash flow “multiple”) are summarized by the parameter g_j . More generally, the growth component is related to any characteristic that scales up prices, relative to cash flows.

Finally, the last term reflects *data*: the more information the average investor has about cash flows, the lower is $\bar{\Sigma}_j$ and therefore, the higher is $PINF_{js}$. This link is what makes *PINF* an informative moment for our purposes. Our empirical strategy involves estimating the growth and volatility components from observables and using them to recover the data component from the observed *PINF*.⁸

2. Estimation of the Structural Model

This section describes how we estimate our structural model to construct our data measure. We describe our sample in detail, as well as construction of variables and moments used for the structural estimation. We also discuss how these estimates relate the corresponding objects in the model.

2.1 Data sample and data adjustments

All data are for the U.S. market, over the period 1962–2016. Stock prices come from CRSP (Center for Research in Security Prices). All accounting variables are from Compustat. We measure prices at the end of March and accounting variables at the end of the previous fiscal year, typically December. This timing convention ensures that market participants have access to the accounting variables that we use as controls. In line with common practice, we exclude firms in the finance industry (SIC code 6).

The equity valuation measure, that is, the empirical counterpart for the price P_{fj1}^* in the model, is market capitalization over total assets, denoted by $M_{f,j,t}^*/A_{f,j,t}^*$. For our cash flow variable, d_{fjs}^* , is proxied using earnings over assets. More precisely, we take earnings before interest and taxes (the *EBIT* variable in Compustat), denoted by $E_{f,j,t}^*$ and divide by current total assets $A_{f,j,t}^*$. Both ratios are winsorized at 1%.

We make a couple of adjustments to the raw data. The first is to deal with inflation, which can create predictability in nominal earnings and prices. This is particularly relevant for periods of high inflation, such as the 1960s and 1970s. Therefore, we adjust all cash flow variables with a gross domestic product

⁸ Note that this *data* component reflects the effect of both information extracted from the price signal and net private data processing. We will show how we can disentangle these different types of data from observable time series.

(GDP) deflator. The second pertains to exiting firms. Our preferred solution is to only consider periods during which a firm has nonmissing information.⁹

Finally, motivated by our focus on measuring stock-specific data, we remove the common (or aggregate) components from both cash flows and prices. To do this, we first construct the analogous “market” variables using total assets, market capitalization and *EBIT* for the universe of S&P 500 firms. Then, separately for each stock in our sample, we project our cash flow and price series for the period 1960–2012 on the corresponding market variables (and a constant) and extract a residual. In what follows, we denote this firm-specific component of prices and cash flows by $\frac{M_{f,j,t}}{A_{f,j,t}}$ and $\frac{E_{f,j,t}}{A_{f,j,t}}$, respectively.

2.2 Variable construction

2.2.1 Size, growth, and volatility. Sorting, measuring, and mapping these variables to the model is critical for our approach. We start by describing our strategy to sort individual stocks into groups. We choose two particular characteristics to construct our groups: size and growth. This choice is motivated by two considerations. First, as we will show in Section 3, the value of data to an investor is closely tied to the overall size of the asset and the growth prospects. Second, these are canonical asset pricing groups, so using them allows us to make contact with the empirical asset pricing literature that examines how large and high-growth stocks differ from their small and low-growth counterparts.¹⁰ At the same time, the reader should not be led into thinking that we are pricing risk factors, as would traditionally be done in that literature. Recall that our price and cash flow variables have been stripped of common factors, leaving only firm-specific components. As such, we are looking at whether firms with these size and growth characteristics have different prevalence of data about their *firm-specific* cash flows.

We group firms into Large and Small, based on whether or not they belong to the 500 largest firms in terms of market capitalization. Next, we classify firms into High-growth and Low-growth based on their book-to-market ratio (defined as the difference between total assets and long-term debt, divided by the firm’s market capitalization). Firms in the top-three deciles of book-to-market we call low-growth firms, while those in the bottom-three deciles are our high-growth firms. Combining these two dimensions yields four groups: Small High-growth, Large High-growth, Small Low-growth, and Large Low-growth. The number of firms in each of these groups for each decade starting with the 1960s is reported in Table B1 in the appendix.

⁹ Our results are also robust if we make cash flows zero when the firm exits or to use a weighted industry cash flow as a proxy, as in Bai, Philippon, and Savov (2016) (along with the delisting price as the equity valuation variable).

¹⁰ We use the term “low-growth,” instead of “value” to distinguish between the asset characteristic from the asset pricing factor. Of course, we could have used other asset pricing factors (e.g., momentum, beta) to group firms as well, but their link to the value of the data about firm-specific factors is less clear.

To understand why the empirical measures of size and growth correspond to the objects \bar{x} and g in the model, we need to consider why these parameters matter and then ask if these empirical measures capture the relevant concerns.

Size matters because it determines the amount of value available to profit from, with good information. In reality, investors can more easily trade large positions on the equity of larger firms. These are more valuable to learn about because investors can make larger trades on such assets to exploit their informational advantage. In the model, greater size means more shares (higher \bar{x}). Constructing the empirical analogue requires specifying what a share means. This is just a normalization: 500 shares worth \$2 each, with a variance of 4, or one with 1,000 shares, worth \$1 each, with a variance of 1, are isomorphic representations. Our empirical notion of a share corresponds to a claim on \$1 worth of the underlying assets. Thus, the number of shares is simply given by the value (in dollars) of the firm's assets. We then measure everything else—prices, cash flows—consistent with this normalization. Specifically, prices and cash flows “per share” are market cap and EBIT, divided by the dollar value of the firm's assets.

Growth (g) matters for data because it scales the earnings-to-valuation ratio. Firms with high g have prices that are a high multiple of earnings and therefore have prices that are very sensitive to earnings news. Growth is a scaling factor. In the data, market-to-book performs a similar function. It scales up the asset's value for a given level of earnings. In both cases, growth increases the loading of prices on cash flows, through the Gordon growth term $\frac{r}{r-g}$. In other words, the same amount of cash flow data affects high-growth firms' prices by more.

The starting point for our approach to measuring data is estimating stock-specific price informativeness as defined in (12) and characterized in (13). Recall that this moment captures the extent to which stock prices in year t reflect cash flows in year $t+s$ and can be estimated from a regression of the latter on the former, along with controls for other observable asset characteristics. Given our interest in long-term trends, we perform this exercise separately for each of the 4 groups in each decade (starting with the 1960s). Specifically, we run the following cross-sectional regression separately for each asset group j and decade:

$$\frac{E_{f,j,t+s}}{A_{f,j,t}} = \alpha_j + \beta_{j,s} \cdot \frac{M_{f,j,t}}{A_{f,j,t}} + \gamma_j \cdot X_{f,j,t} + \epsilon_{f,j,t+s} \quad (14)$$

where $E_{f,j,t+s}/A_{f,j,t}$ is the cash flow (*EBIT*) of firm f in group j in year $t+s$, scaled by its total assets in year t ; $\log(M_{f,j,t}/A_{f,j,t})$ is market capitalization scaled by total assets; and $X_{f,j,t}$ are a set of firm-level controls, namely, past earnings and industry fixed effects, meant to capture publicly available information. We use $s=3$ in our estimation.

To obtain the measure in (12), we scale the coefficient $\beta_{j,s}$ by the variability of the regressor:¹¹

$$PINF_{js} = \beta_{js} \cdot \sigma_j^{M/A}, \quad (15)$$

where $\sigma_j^{M/A}$ denotes the cross-sectional standard deviation of $\frac{M_{f,j,t}}{A_{f,j,t}}$ (conditional on controls). This strategy and the measure $PINF_{js}$ is very closely related to the one in Bai, Philippon, and Savov (2016).¹² From our perspective, it is a convenient starting point for recovering the object we are ultimately interested in, namely, the extent of data processed about firm-specific factors.

3. Results

Next, we employ this framework to measure cross-sectional and time series patterns of data in the market. We further separate this information into cross-asset differences in the efficiency with which market aggregates net private data, that is, investor data above and beyond what could possibly be extracted from prices.

Equation (13) shows that we need to remove the effects of volatility and growth from the estimate of informativeness in order to isolate the data component. For this, we need quantitative estimates of these two components for each group and decade. The volatility component is related to the variability of the unpredictable innovation in cash flows and the (conditional) standard deviation of prices. These are estimated by projecting our cash flow and price measures on a set of controls and calculating the standard deviation of the residuals (again, separately for each group and decade). The bottom panel of Table 1 reports the resultant estimates for the variance of the innovation to cash flows (Σ_{jd}). Dividing this by the (conditional) standard deviation of prices yields the volatility component in (13).

Next, we turn to the estimation of the growth component. Recall that this term arises because growth rates influence the factor by which earnings are scaled in the equilibrium pricing Equation (7): in other words, the growth factor $\frac{r}{r-g_j}$ converts per-period cash flows to the same units as price. It is related to the price-earnings ratio, though the latter will also pick up effects of informational frictions and risk premiums.

We estimate growth rates (by group and decade) by running regressions of cash flows on their lagged values. The resultant autoregressive coefficients map directly onto g_j and are reported in the top panel of Table 1. As we would expect, high-growth firms generally have higher growth rates (relative to their

¹¹ We use the absolute value of the estimated price informativeness, since the theory cannot reconcile negative estimates (this only matters for a couple of observations and does not affect conclusions about longer-term trends).

¹² There are some important differences, both conceptual and measurement related. See Section 3.2 and Appendix E for more details.

Table 1
Estimated cash flow parameters: Persistence/growth g_j , variance of innovation Σ_{jd}

	1960s	1970s	1980s	1990s	2000s	2010s
Persistence g_j :						
Small/high growth	0.830	0.877	0.702	0.725	0.740	0.741
Large/high growth	0.988	0.981	0.954	0.949	0.917	0.912
Small/ low growth	0.829	0.697	0.538	0.572	0.669	0.636
Large/ low growth	0.901	0.865	0.853	0.813	0.828	0.851
Variance of Innovations Σ_{jd} :						
Small/high growth	0.005	0.007	0.019	0.022	0.017	0.013
Large/high growth	0.002	0.002	0.002	0.003	0.003	0.003
Small/ low growth	0.002	0.004	0.009	0.008	0.009	0.006
Large/ low growth	0.002	0.001	0.001	0.001	0.002	0.001

Persistence g_j is estimated by running regressions of cash flows on their lagged values, as specified in Equation (1). Σ_{jd} is estimated as the variance of residuals from a projection of cash flows on controls.

counterparts in the corresponding low-growth category). Assuming a riskless interest rate of 2.5% ($r = 1.025$), these estimates directly yield $\frac{g_j^s}{r - g_j}$, the growth component.¹³

3.1 Data divergence

The estimates of PINF, along with the corresponding growth and volatility components in (13), allow us to back out the information component $1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}}$, our measure of data. Specifically, for each decade group, we divide the estimated PINF for that decade group by the corresponding growth and volatility components, as defined in (13), to back out the underlying data term:

$$1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} = \frac{PINF_{j,s}}{\frac{g_j^s}{r - g_j} \frac{\Sigma_{jd}}{Std(P_{fj1})}}. \tag{16}$$

Equation (16) clearly shows how the three key empirical objects that we measure, namely, price informativeness, growth, and volatility, drive our estimate of data. Holding growth and volatility fixed, a higher *PINF* is a sign of more data processing. And for a given *PINF* estimate, higher growth or greater volatility of cash flows (relative to prices) correspond to less data processing.

Our decade-by-group estimates for *PINF* and its components¹⁴ are reported in Table 2. We plot these estimates along with fitted linear trend lines for each series in Figure 1. Critically, the top-left panel shows that *PINF* has trended

¹³ In our baseline analysis, we use $r = 1.025$ for the entire sample. In Appendix B, specifically in Figure B1, we relax this assumption and show that our results are robust to using decade-specific values for interest rates.

¹⁴ For one decade group, the right-hand side of (16) produced an estimate larger than one, which would be inconsistent with the structural model. We therefore top-coded those estimates using a bound of 0.95. This adjustment made only a negligible difference to the overall trends.

Table 2
Stock-specific price informativeness and its components

	1960s	1970s	1980s	1990s	2000s	2010s
PINF						
Small/high growth	0.012	0.003	0.003	0.008	0.013	0.007
Large/high growth	0.015	0.014	0.018	0.011	0.037	0.023
Small/ low growth	0.003	0.007	0.003	0.002	0.011	0.004
Large/ low growth	0.003	0.006	0.002	0.003	0.014	0.001
Volatility						
Small/high growth	0.005	0.008	0.012	0.013	0.011	0.010
Large/high growth	0.002	0.002	0.003	0.002	0.002	0.003
Small/ low growth	0.016	0.054	0.064	0.044	0.037	0.034
Large/ low growth	0.012	0.017	0.012	0.011	0.013	0.008
Growth						
Small/high growth	2.94	4.56	1.07	1.27	1.42	1.44
Large/high growth	25.87	21.34	12.33	11.31	7.17	6.69
Small/ low growth	2.91	1.03	0.32	0.41	0.84	0.66
Large/ low growth	5.91	4.06	3.60	2.54	2.88	3.53
Data						
Small/high growth	0.84	0.08	0.22	0.50	0.80	0.50
Large/high growth	0.36	0.40	0.57	0.44	0.95	0.95
Small/ low growth	0.06	0.12	0.12	0.10	0.36	0.17
Large/ low growth	0.04	0.09	0.04	0.10	0.38	0.03

The table reports structurally estimated values of the various terms in Equation (13) using cash flow parameters in Table 1. The left-hand side is estimated using Equations (14) and (15).

up for the Large/High-growth group, much faster than for all other groups. The top-right panel reveals that changes in data played a central role in the divergence.

The remaining panels in Figure 1 show the trends in other components. In particular, the growth component (bottom-right panel) highlights why it is important to distinguish between data and a measure like price informativeness.¹⁵ Growth declines most dramatically for large/high-growth assets. By itself, this trend should have reduced the informativeness of (the stock-specific components of) those assets. Had this change been larger, we might have found *PINF* and data moving in opposite directions. Instead, the rise in data about large/high-growth firms was sufficiently large that it overwhelmed the effect of declining growth on informativeness.¹⁶

Next, we explore where the trends in data came from. Was the firm specific data information mined from public prices, or was it extracted from other sources? To answer this question, we decompose overall information, $\bar{\Sigma}_j^{-1}$, into its components as in (6). Specifically, the prior or unconditional precision (Σ_{jd}^{-1}), the information content of the price signal (Σ_{jp}^{-1}) and the net private data (K_j). To estimate the second component, namely, the information conveyed by the price signal, we run (14) with $s=0$ and calculate the variance of the residuals,

¹⁵ The decline in g_j for large/high-growth firms is consistent with Gschwandtner (2012), who also finds a long-run decline in the persistence of firms' profit. This could reflect, for example, an increase in competition because of globalization.

¹⁶ In Table B2 in Appendix B, we show that the differences across groups in the trends in data are statistically significant.

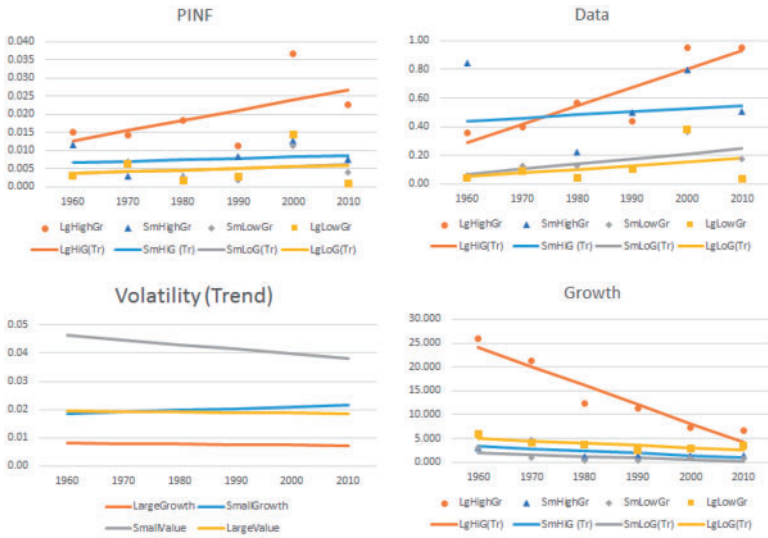


Figure 1
Data divergence: Trends in data and other components of stock-specific price informativeness
Graphical representation of Table 2. For each component, the dots represent the estimates reported in Table 2, while the corresponding lines represent the (linear) trend.

denoted by $Var(e_{jj})$. Appendix A.3 derives the following mapping between Σ_{jp} and $Var(e_{jj})$:

$$\Sigma_{jp} = \frac{Var(e_{jj}) \cdot \Sigma_{jd}}{\Sigma_{jd} - Var(e_{jj})}. \quad (17)$$

Substituting the resultant estimates of Σ_{jp} , along with the overall market information $\bar{\Sigma}_j^{-1}$ and the prior precision, Σ_{jd}^{-1} , into (6) yields the net private data for each group-decade, K_j :

$$K_j = \bar{\Sigma}_j^{-1} - \Sigma_{jd}^{-1} - \Sigma_{jp}^{-1}. \quad (18)$$

Table 3 presents the estimates for price information Σ_{jp}^{-1} and net private data K_j , by group and decade. Figure 2 plots the associated fitted trend lines. They show a generally declining trend in firm-specific market information across all groups. The trends in total data, most notably the rise for Large/High-growth stocks, can be attributed mostly to changes in net private rather than price information.

Note from Table 3 that the estimates for net private data K_j are negative in some cases, particularly in the early part of the sample for low-growth stocks. This happens when the PINF (or more precisely, the price-earnings covariance) is less than what it would be if all investors were learning the maximum possible from market prices. In other words, this pattern suggests that the average investor may not fully process all the information contained in

Table 3
Sources of information

	1960s	1970s	1980s	1990s	2000s	2010s
Price information, Σ_{jp}^{-1} :						
Small/high growth	10	1	0	1	1	0
Large/high growth	80	18	82	41	55	16
Small/ low growth	77	20	3	0	3	0
Large/ low growth	130	105	72	41	8	1
Net private information, K_j :						
Small/high growth	1013	12	15	45	236	75
Large/high growth	269	362	451	193	7151	5868
Small/ low growth	-49	13	14	13	59	33
Large/ low growth	-106	-33	-40	45	258	30

Price information and net private data are estimated using Equations (17) and (18), respectively.

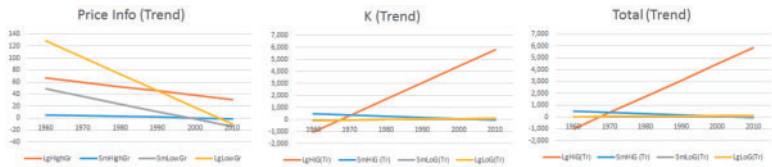


Figure 2
Rising large/high-growth firm data come from net private information

Graphical representation of the trends in the estimates reported in Table 3. For each component, the plot shows the linear trend fitted to the estimates from the table. Total is the sum of Price Information (Σ_{jp}^{-1}) and Net Private Information (K). This total is the same as Data, plotted in Figure 1.

prices (e.g., because learning from prices is also costly). In such a scenario, our approach to decomposing total data would overestimate price information Σ_{jp}^{-1} , or equivalently, underestimate K_j . However, since price information accounts for only a small fraction of total information, this source of mismeasurement is small, relative to the trends in data processing.

3.1.1 Dating the data revolution. Table 3 also tells us when financial markets started to embrace big data: net private data rose sharply during the 2000s for all groups. Investors in all four types of assets more than quadruple their private precision between the 1990s and 2000s. This is the same time as the widespread adoption of information technology in the financial sector (Abis 2020) and is consistent with a rapid advance in data technology in the last two decades. But, more interestingly, this rise was the most stark for the large high-growth firms: in other words, the data revolution disproportionately favored learning about large/high-growth firms, contributing significantly to the trend of data divergence.

3.1.2 Adjusting data for market power. We know that market power can effect price informativeness, but how does it affect our measure of data? From

?, we know that incorporating market power involves replacing the conditional variance $V[f_t|\mathcal{I}_i]$ with $V[f_t|\mathcal{I}_i] + \lambda/\rho$, where λ is Kyle's lambda, the price impact of a unit of demand and ρ is absolute risk aversion.

What this means for measurement is that, if we are ignoring investor market power, we then are overestimating the conditional variance. Conversely, we are underestimating data precision. So, our estimates might be considered a lower bound on the size of the data stock. However, the measure of $V[f_t|\mathcal{I}_i] + \lambda/\rho$ is useful, by itself. This sum is the object that appears in the marginal value of data when investors have market power.

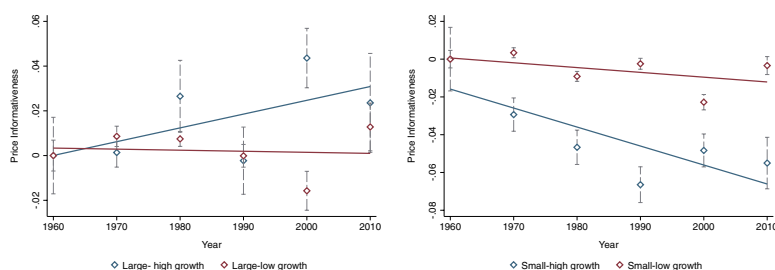
3.2 Relating data to other information measures

There are a number of measures for the information embedded in prices in the literature. In this subsection, we clarify how our *data* measure is different from them.

3.2.1 Price informativeness. As we discussed earlier, *PINF*, which measures stock-specific price informativeness is an input in our measurement strategy and is closely related to the one in Bai, Philippon, and Savov (2016), is conceptually different from our data measure. *PINF* measures the extent to which (the stock-specific components of) prices and future cash flows covary, which is also affected by growth and volatility effects in addition to data. We develop a tool designed to isolate the latter and show that it has diverged over the last few decades.

Our analysis also differs in its focus on firm-specific factors. We remove aggregate/common components from both cash flows and prices, while Bai, Philippon, and Savov (2016) work with unadjusted cash flows and so pick up informativeness of prices with respect to both common and stock-specific factors. There are a couple of other measurement differences as well. First, we work with $\frac{MktVal}{Assets}$ in levels, rather than logs, to be consistent with our structural model. Furthermore, the price informativeness measure in Bai, Philippon, and Savov (2016) is obtained by scaling the regression coefficient of the current price by the *unconditional* standard deviation of prices, while our structural framework suggests scaling by the standard deviation of prices *conditional* on the controls. These adjustments tighten the connection to the structural model and affect magnitudes but, as we will show below, do not significantly change the overall trends.

Figure 3 plots the trend in the price informativeness measure of Bai, Philippon, and Savov (2016), estimated decade-by-decade for four subsamples of firms: Large/High-growth, Large/Low-growth, Small/High-growth and Small/Low-growth. It shows that informativeness has increased for the Large/High-growth group, but declined for the others. While Bai, Philippon,

**Figure 3****Price informativeness by firm size and growth**

The diamonds represent the estimated price informativeness defined as in Bai, Philippon, and Savov (2016), along with 95% confidence intervals. For the detailed specification, see Equation (E.2) in Appendix E. Large firms are the 500 largest firms by market capitalization, and small denotes the rest. Firms in the bottom (top) 30% are labeled high-growth (low-growth) firms. The lines represent fitted trend lines.

and Savov (2016) also note the divergence between firms in and out of the S&P 500, we show that this is a size effect, not an index inclusion effect.¹⁷

3.2.2 Comovement, R^2 , and synchronicity. Many papers have applied comovement, synchronicity or R^2 approaches to measuring stock market informativeness across countries (Durnev, Morck, and Yeung 2004; Edmans, Jayaraman, and Schneemeier 2017). These measures are valuable tools for cross-country analysis of price movements, but are not appropriate for measuring the precision of data, about one type of firm versus another. For example, asset comovement (R^2) could be high because of aggregate information is precise, causing many assets to move with that aggregate information, or because stock-specific information is imprecise. Any mapping to data precision requires decomposing aggregate and stock-specific data, which in turn, requires an independent measure of one or the other. Our approach explicitly constructs that measure and uses the structure of the model to back out data processing.

Furthermore, an R^2 measure shares many of the same interpretation problems of the price informativeness measure. To see why, note from the pricing equation (7), if the aggregate cash flow shock $\bar{\epsilon}$ is observed, the R^2 is the explained sum of squares $B^2 \text{var}(\bar{\epsilon})$, divided by the total sum of squares, $B^2 \text{var}(\bar{\epsilon}) + C^2 \text{var}(\epsilon) + D^2 \text{var}(\tilde{x})$. Just like PINF, these quantities depend on the coefficients, like B and C , which are affected by the amount of data, but also are contaminated by volatility and scaling terms like growth. For example, Brogaard et al. (2018) argue that stock return comovement, as measured by

¹⁷ Appendix E.2 shows that the informativeness of stocks currently in the S&P 500 is similar to non-S&P 500 stocks with similar characteristics. Furthermore, price informativeness trends consistently over size deciles. These results suggest that differences in asset characteristics, rather than inclusion in S&P 500 *per se*, is the source of the divergence.

R^2 , has increased significantly over time, because idiosyncratic price noise declined.

3.2.3 Absolute price informativeness. Davila and Parlato (2016b) propose an alternative measure of “absolute price informativeness,” which captures the ability of asset prices to aggregate dispersed information. Their measure is the precision of an unbiased signal of the current cash flow innovation, constructed from prices. In our setting, this corresponds to Σ_{jp}^{-1} , what we call “price information” in Table 3 and Figure 2.

As Figure 2 shows, absolute price informativeness declines across all four asset groups. Thus, despite *more* net private data processed by investors about stock-specific characteristics (higher K_j) prices actually became *less* accurate as signals (the Davila-Parlato notion of price informativeness). This difference arises because the noise component of prices (from the $C_j \tilde{x}_j$) grew over time and overwhelmed the rising covariance with fundamentals.

This finding differs from Davila and Parlato (2016b) primarily because we strip out the aggregate component of prices and cash flows, while they do not. If we redo their exercise with raw prices and earnings, price information does show a rising trend. These results suggest that prices may be getting better at aggregating market information, but are becoming less clear signals about firm-specific cash flow risk.

3.2.4 Other proxies for data. Many papers explore proxies for information, including news counts, analyst coverage, advertising, or social media text (see the examples cited below). Of course, these proxies are useful for qualitative validation and do not obviate the need for a quantitative measure like ours. Moreover, while these proxies document interesting cross-sectional patterns, to the best of our knowledge, none of them focuses on how qualitative patterns have changed over time, especially over the horizons we are interested in.

Coverage by equity analysts on Wall Street is a natural proxy for information processing. Hong, Lim, and Stein (2000) and Guo and Mota (Forthcoming) analyze determinants of coverage, but do not discuss time trends. In Appendix D, we estimate time trends in analyst coverage (using the I/B/E/S database of analyst forecasts) and show that there was a sharp increase in the relative coverage of high-growth firms during the 2000s and 2010s. This is particularly striking for large firms and the timing of this increase lines up quite well with the results of our structural approach.

Of course, it is worth noting that analyst coverage is likely a rather crude measure of data precision. For one, it doesn’t capture variation in quality of data processing, both in the cross-section and over time. An analyst might be reporting mostly redundant or low-quality information that does little to reduce investor uncertainty (in fact, to the extent it disagrees with other analysts’ forecasts, it might even seed uncertainty). Moreover, analyst coverage also does not capture data processing done in-house by investors (e.g., hedge funds),

which has arguably displaced work traditionally done by sell-side analysts over time. This evidence, albeit reassuring and suggestive, neither displaces the need for a data precision measure nor reveals the source of the divergent data trends.

Firms also use advertising to convey information to outsiders. Chemmanur and Yan (2019) examine the effect of such advertising on stock prices and find that the effects are smallest for large/high-growth firms. Our model suggests that new information, such as that contained in an ad, is likely to have small effects when the existing information is already of high quality. In other words, one explanation for findings in Chemmanur and Yan (2019) is that data on large high-growth firms is relatively abundant, consistent with our results. Note that similar to the literature on analyst coverage, this paper also focuses on the cross-section and not changes over time. Nevertheless, the fact that the cross-sectional evidence is broadly consistent with our story is a reassuring finding.

4. Why Did Large High-Growth Firm Data Become More Abundant?

Our results show that, while asset characteristics did change over this period, divergence in the price informativeness for Large/High-growth firms came predominantly from data divergence. This raises an obvious question: why did so many investors process increasing amounts of data about large high-growth stocks and not about other assets?

One possibility is that data choices changed over time because the cost of data changed. For this to explain our findings, the cost of large high-growth firm data must be falling, relative to the cost of data about other firms. Given that we have no direct evidence to support or quantify this channel, we focus on the relative benefit of data on large high-growth firms. Here, our structural framework can help us talk about how observed changes in asset characteristics should change the value of data and through that, data choices. This is the approach we take: we abstract from differences in costs, use the estimates from the structural model to see whether the model-implied benefit of data has changed in a manner consistent with observed patterns in data choices across groups. This does not rule out—and in fact, is complementary to—the possibility of changes in relative costs.

One might be tempted to look at equilibrium *marginal* values for this purpose. However, they are not very useful in predicting the *amount* of data allocated to different assets. This is because equilibrium forces push to equate marginal values across assets. In other words, in equilibrium, agents will process different amounts of data for different assets up to a point at which the marginal value of additional data processing is the same. We are interested in explaining how much data are processed about a particular asset; the equilibrium marginal value cannot tell us that. The same logic that Berk and Green (2004) applied to mutual fund flows also applies to data flows: equilibrium forces should equalize marginal returns.

If the equilibrium marginal value does not reliably explain the amount of data processing, what does? One candidate is *the initial value of data*, defined as the value of the first increment of precision, that is, the marginal utility gain from a unit of data in a hypothetical world where no one else processed any data on that asset. The basic idea is that assets with the highest initial value will see the most amount of data processed (even if all assets have the same marginal value in equilibrium).¹⁸ If the assets for which data processing is high also have high initial values of information, this could explain the data divergence we see in the previous section.

We use the model to estimate the initial value of one unit of processed data (one precision unit) about each asset type, in each decade. We find that the value of learning about large firms rose substantially over this period, both in absolute terms and relative to small firms. The divergence in data value was driven by the increase in large firms' relative size. This surge in the relative size of large firms has been documented by the firm dynamics literature (e.g., Davis and Haltiwanger 2015). The source of this divergence is the subject of an active debate in the macroeconomics and IO literatures.

4.1 Derivation of initial value of information

To arrive at the value of information, we compute the ex ante expected utility and determine its sensitivity to information choice. Ex ante expected utility of investor i from assets in group j is given by

$$\mathbb{E}[U_j^i] = \frac{1}{2} \mathbb{E}[(\Pi_j^i)^2] \left(\frac{r}{r - g_j} \right)^{-2} (\Sigma_j^i)^{-1} \quad \text{where} \quad \Pi_j^i \equiv \mathbb{E}[V_j - P_j r | \mathcal{I}^i]. \quad (19)$$

(Π_j^i) is the interim (i.e., conditional on a data set \mathcal{I}^i) expected profit per share of asset j , and $(\Sigma_j^i)^{-1}$ is investor i 's posterior precision about cash flows. This form of expected utility arises in a large class of noisy rational expectations models. Intuitively, investor i 's interim profits are $q_j^i \Pi_j^i$. The optimal asset demand q_j^i is proportional to $\text{Var}[V | \mathcal{I}^i]^{-1} \Pi_j^i$ where $\text{Var}[V | \mathcal{I}^i]^{-1} = \left(\frac{r}{r - g_j} \right)^{-2} (\Sigma_j^i)^{-1}$.

Equation (19) directly shows that the marginal utility of a unit increase in the investor's posterior precision is $\frac{1}{2} \mathbb{E}[(\Pi_j^i)^2] \left(\frac{r}{r - g_j} \right)^{-2}$. This is the marginal value of data. Data are more valuable when profits are expected to be high (in

¹⁸ This concept is related to what is sometimes referred to as a water-filling equilibrium in the information choice literature. In equilibrium, agents sequentially choose risk factors to learn about: learning about a risk is like filling its "bucket" with water. Once sufficiently full, investors move on to filling the next deepest bucket. Our value of information can be thought of the depth of each bucket, before being filled with water. At the optimum, all buckets will be filled to the same level (equal marginal value), but the deepest buckets will hold the most amount of water.

absolute value)¹⁹ and/or more volatile because that makes the expected value of the squared profit high.

Next, we compute the unconditional expected profit per share.²⁰

$$\mathbb{E}[\Pi_j^i] = \bar{\rho} \left(\frac{r}{r - g_j} \right)^2 \bar{\Sigma}_j \bar{x}_j. \quad (20)$$

Thus, the expected profit per share is the product of the total amount of asset j risk borne by the average investor, scaled by aggregate risk aversion $\bar{\rho}$. Faster growth, or equivalently, a higher valuation-to-cash-flow ratio (higher $\frac{r}{r - g_j}$) means greater uncertainty about the discounted values of the entire cash flow stream, for a given level of uncertainty about current cash flows ($\bar{\Sigma}_j$). Similarly, larger supply (higher \bar{x}_j) implies more risk exposure for the average investor's portfolio and therefore, a larger compensation in the form of expected profits. In other words, it is more valuable to learn about large, fast-growing firms with greater uncertainty.

To compute the *initial* value of data, we simply replace the equilibrium information level $\bar{\Sigma}_j$ with its value before any data are processed, the prior variance Σ_{jd} in (19). Then, compute the partial derivative with respect to $(\Sigma_j^i)^{-1}$. This is what we call the *initial value of information* (VI_j):

$$VI_j = \frac{1}{2} \left[\bar{\rho}^2 \left(\frac{r}{r - g_j} \right)^2 \Sigma_{jd}^2 \bar{x}_j^2 \right] + \frac{1}{2} \Sigma_{jd} \quad (21)$$

The first term in (21) is related to the mean of the expected profit per share of asset j from (20). As we saw earlier, higher growth (g_j), larger size (\bar{x}_j) and more uncertainty (Σ_{jd}) all raise VI_j , making information about the asset's cash flows more valuable. Moreover, these factors enter multiplicatively and therefore, amplify each other. This interaction makes Large/High-growth firms valuable for many investors to learn about.

The second term in (21) stems from the variance of expected profits per share. Quantitatively, however, this term is dominated by the first term, because $\frac{r}{r - g_j}$ and \bar{x}_j are both large, relative to other terms. In other words, most of the variation in the value of the information, both in the cross-section and over time, comes from changes in the size and scale of profitable trading opportunities.

4.2 Estimation of initial value of information

We construct a time series for the value of information (VI_j), for each of the four asset groups by decade. Computing VI_j requires parameters already

¹⁹ High negative expected profits are also valuable, because they present profitable shorting opportunities. The $1/2$ in Equation (19) comes from subtracting a variance term in the formula for the mean of a lognormal variable.

²⁰ Note $\mathbb{E}[(\Pi_j^i)^2] = (\mathbb{E}[\Pi_j^i])^2 + \text{Var}(\Pi_j^i)$. See Appendix C for the derivation of $\text{Var}[\Pi_j^i]$ and other details.

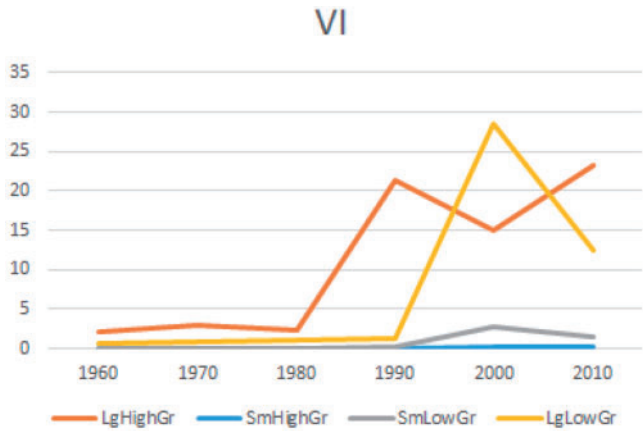


Figure 4
The initial value of information, by asset class, over time
The initial value of information VI_j is defined in (21). The estimates for g_j and Σ_{jd} are reported in Table 1, while the estimates for \bar{x}_j are reported in Table B1. $r=1.025$ and $\bar{\rho}=0.02$.

estimated in Section 2, as well as risk aversion $\bar{\rho}$ and the asset supply (\bar{x}_j). To estimate total supply, we first calculate the average (book) value of assets of firms in group j by decade. We then project firm-level assets on total assets of the S&P 500 and estimate the fraction of the variance that is unexplained by the regressor (i.e., $1-R^2$). The value of assets associated with stock-specific component is then obtained by multiplying this factor (the average value for each group) and the average book value of assets (also by group, reported in Table B1 in the appendix). Finally, we assume the risk aversion coefficient is $\bar{\rho}=0.02$.

The resultant estimates in Figure 4 offer a simple explanation for why so much data has been processed for large firms, especially large high-growth firms. Information about such firms is more valuable. Both size and growth increase the value of information, which is also amplified by their interaction. The combination of being large and growing quickly makes a firm a desirable target for data analysis. In the figure, the value of information for small high-growth and small low-growth stocks is very close to zero, orders of magnitude lower than the value of the large firms' data.

The time series for VI_j in Figure 4 shows a dramatic rise in the value of large firms' information during the 1990s and 2000s. These patterns are driven almost entirely by movements in the first term in (21). Why did this component rise so sharply and then fall? The increase can be traced to the rise in their size (i.e., the amount of assets, \bar{x}_j in the model): in other words, large firms grew larger (both in absolute and in relative terms) during the 1990s and 2000s, raising expected profits per share and making data about them more valuable.

The value of large low-growth firms' information surpasses that of large high-growth firms for one decade in our sample. This was likely the combined result of a decrease in the growth prospects of large high-growth firms and a rise in the relative size of large low-growth firms. One possibility is that these changes in firms' characteristics was unexpected. If data processing can be frictionlessly reallocated, one would expect a quick reaction to the surprise change in growth and size. But, in reality, research expertise takes time to build: time to hire personnel, and time for them to develop the necessary knowledge. As a result, it is quite likely that, much like physical capital, information processing is slow to adjust. A full exploration of this possibility is a question for another paper.

5. Conclusions

Financial services are increasingly centered on data processing. Making optimal data choices and valuing data requires knowing the precision of other market participants' forecasting data. We develop a tool to measure this data precision. Our tool can be applied in many possible ways to various groupings of assets.

Since our framework tells us that size and growth make data valuable, we use our tool to measure data for firms sorted by size and growth. We find data divergence: investors seem to be processing more and more data about large high-growth assets, but not about others.

To explore why data processing might diverge, we use the estimated structural model to impute a value of data. We find that the value of large high-growth firm data has increased, primarily because these firms grew larger. Larger firms are more valuable to learn about, particularly if they are also expected to grow faster. While our tool has uncovered a new fact and suggested a logical explanation for it, there surely will be many reasons to want to measure data along other dimensions, as we continue to learn more about the financial data economy.

Appendix

A. Structural Framework: Derivations

A.1 Proof of Proposition 1

Solving for the equilibrium follows a standard guess-and-verify procedure, widely used in the noisy rational expectations equilibrium (REE) literature. First, we express total demand for each asset j , as a function of price (P_{j1}), and equate it with total supply ($\bar{x} + \bar{x}_j$). Then, we match coefficients on both sides of this market clearing condition to obtain a system of equations in A_j, B_j, C_j . Specifically, all constant terms are equated to A_j ; terms that multiply ϵ_{j1} are equated to B_j , and, finally, those multiplying \bar{x}_j must equal C_j . Simplifying that system of equations yields the stated result.

A.2 Decomposing Price Informativeness: Derivation of Equation (13)

$$PINF_{j,s} = \frac{Cov(d_{fjs}^*, P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)}{StdDev(P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)} = g_j^s \frac{Cov(d_{fj1}^*, P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)}{StdDev(P_{fj1}^* | d_{fj0}^*, \bar{\epsilon}_1)} \quad (A.1)$$

$$= g_j^s \frac{Cov(\epsilon_{fj1}, P_{fj1})}{StdDev(P_{fj1})} = \frac{g_j^s}{r} \frac{B_j \Sigma_{jd}}{StdDev(P_{fj1})} \quad (A.2)$$

$$= \frac{\Sigma_{jd}}{StdDev(P_{fj1})} \frac{g_j^s}{r - g_j} \left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right), \quad (A.3)$$

where the last line uses the expression for B_j from (9).

A.3 Estimating Σ_{jp} : Derivation of Equation (17)

The stock-specific components, that is, residuals after conditioning on $(d_{fj0}^*, \bar{\epsilon}_1)$, are given by

$$d_{fj1} = \epsilon_{fj1} \quad (A.4)$$

$$P_{fj1} = \tilde{A}_j + \frac{B_j}{r} \epsilon_{fj1} + \frac{C_j}{r} \tilde{x}_{fj}, \quad (A.5)$$

where $\tilde{A}_j = -\bar{\rho} \left(\frac{r}{r - g_j} \right)^2 \bar{\Sigma}_j \tilde{x}_j$. The coefficients from regressing d_{fj1} on P_{fj1} and a constant are

$$\hat{\beta}_j = \frac{Cov(\epsilon_{fj1}, P_{fj1})}{Var(P_{fj1})} = \frac{r B_j \Sigma_{jd}}{B_j^2 \Sigma_{jd} + C_j^2 \Sigma_{jx}},$$

$$\alpha_j = \mathbb{E}(\epsilon_{fj1}) - \hat{\beta}_j \mathbb{E}(\tilde{A}_j + (B_j/r) \epsilon_{fj1} + (C_j/r) \tilde{x}_j) = -\hat{\beta}_j \tilde{A}_j,$$

where we use $\mathbb{E}[\epsilon_{fj}] = \mathbb{E}[\tilde{x}_{fj}] = 0$. The estimated residuals and their variance are

$$\begin{aligned} e_{fj} &= \epsilon_{fj1} - \alpha_j - \hat{\beta}_j \left(\tilde{A}_j + \frac{B_j}{r} \epsilon_{fj1} + \frac{C_j}{r} \tilde{x}_{fj} \right) \\ &= \left(1 - \hat{\beta}_j \frac{B_j}{r} \right) \epsilon_{fj1} - \hat{\beta}_j \frac{C_j}{r} \tilde{x}_{fj}, \\ &= \left(1 - \frac{B_j \Sigma_{jd}}{B_j^2 \Sigma_{jd} + C_j^2 \Sigma_{jx}} B_j \right) \epsilon_{fj1} - \left(\frac{B_j \Sigma_{jd}}{B_j^2 \Sigma_{jd} + C_j^2 \Sigma_{jx}} \right) C_j \tilde{x}_{fj}, \\ &= \left(\frac{C_j^2 \Sigma_{jx}}{B_j^2 \Sigma_{jd} + C_j^2 \Sigma_{jx}} \right) \epsilon_{fj1} - \left(\frac{B_j^2 \Sigma_{jd}}{B_j^2 \Sigma_{jd} + C_j^2 \Sigma_{jx}} \right) \frac{C_j}{B_j} \tilde{x}_{fj}, \\ &= \left(\frac{\frac{C_j^2}{B_j^2} \Sigma_{jx}}{\Sigma_{jd} + \frac{C_j^2}{B_j^2} \Sigma_{jx}} \right) \epsilon_{fj1} - \left(\frac{\Sigma_{jd}}{\Sigma_{jd} + \frac{C_j^2}{B_j^2} \Sigma_{jx}} \right) \frac{C_j}{B_j} \tilde{x}_{fj}, \\ \Rightarrow Var(e_{fj}) &= \left(\frac{\frac{C_j^2}{B_j^2} \Sigma_{jx}}{\Sigma_{jd} + \frac{C_j^2}{B_j^2} \Sigma_{jx}} \right)^2 \Sigma_{jd} + \left(\frac{\Sigma_{jd}}{\Sigma_{jd} + \frac{C_j^2}{B_j^2} \Sigma_{jx}} \right)^2 \frac{C_j^2}{B_j^2} \Sigma_{jx}. \end{aligned} \quad (A.6)$$

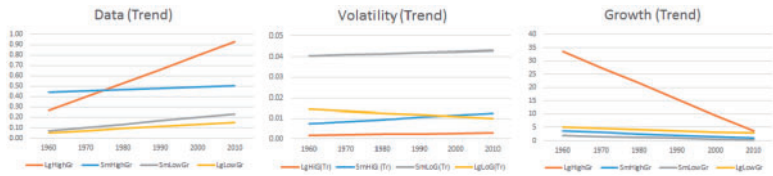


Figure B1
Time variation in riskless rate

The plots depict a linear trend fitted to the structural estimates of the components of PINF as described in (13) and decade-specific interest rates. For details of how the interest rates r are estimated, see the main text.

Noting that $\Sigma_{jp} = \frac{c_j^2}{B_j^2} \Sigma_{jx}$, we can write (A.6) more succinctly as

$$Var(e_{fj}) = \left(\frac{\Sigma_{jp}}{\Sigma_{jd} + \Sigma_{jp}} \right)^2 \Sigma_{jd} + \left(\frac{\Sigma_{jd}}{\Sigma_{jd} + \Sigma_{jp}} \right)^2 \Sigma_{jp} = \frac{\Sigma_{jp} \Sigma_{jd}}{\Sigma_{jd} + \Sigma_{jp}}. \tag{A.7}$$

Solving (A.7) for Σ_{jp} yields the expression in (17).

B. Structural Estimation: Details and Additional Results

B.1 Sample size

Table B1 lists the number of firms and average value of assets for the firms in our sample, separately for each decade and each asset group.

B.2 Time-varying interest and growth rates

In our baseline estimation, we assumed a constant $r = 1.025$ over time. In this subsection, we show that this is not a critical assumption. In particular, we compute the actual average real interest rate for each decade (defined as the difference between 1-year nominal Treasury yield from the Federal Reserve Board’s H15 release and realized inflation over the subsequent year, computed using the PCE Price index) and use that series to reestimate the growth and information components of price informativeness (note that the volatility component remains unaffected). Figure B1 plots the estimated trends for all three components and looks very similar to the baseline results in Figure 1.

B.3 Trends in price informativeness and data

Table B2 shows the point estimates (and associated standard errors) for trends, both in absolute terms and in differences between trends in the Large/High-growth group and the other groups.

Table B1
Number of firms and total assets by decade and type

	1960s	1970s	1980s	1990s	2000s	2010s
Number of firms						
Small high growth	1,699	4,739	7,224	9,253	6,444	3,505
Large high growth	1,696	4,229	6,270	7,963	5,662	3,327
Small low growth	1,734	4,664	7,229	9,153	6,382	3,472
Large low growth	1,653	4,040	6,146	7,742	5,534	3,272
Average assets (\$ millions)						
Small high growth	125	173	109	175	410	599
Large high growth	2,697	3,510	3,521	8,661	12,928	13,802
Small low growth	517	565	852	2,140	4,478	5,398
Large low growth	6,129	11,592	15,726	22,003	52,550	61,588

Table B2
Trends in price informativeness and data

	Trend	SE	Diff in trend	SE
PINF				
Small/high growth	0.0004	(0.0011)	0.0024	(0.0023)
Large/high growth	0.0028	(0.0020)	-	-
Small/ low growth	0.0005	(0.0009)	0.0023	(0.0022)
Large/ low growth	0.0004	(0.0013)	0.0024	(0.0024)
Data				
Small/high growth	0.0207	(0.0800)	0.1072	(0.0873)
Large/high growth	0.1279**	(0.0348)	-	-
Small/ low growth	0.0356	(0.0222)	0.0922**	(0.0413)
Large/ low growth	0.0255	(0.0334)	0.1024**	(0.0482)
Net private data				
Small/high growth	-113.8	(87.6)	1,488.2**	(527.0)
Large/high growth	1,374.4**	(519.7)	-	-
Small/ low growth	15.7**	(5.4)	1,358.8**	(519.7)
Large/ low growth	46.8**	(24.3)	1,327.6**	(520.2)

The “Trend” column reports the slope of fitted trendline for PINF, and the “Data” and “Net” rows report private data from Figures 1 and 2, where the time variable is given by $t = \{1, 2, \dots, 6\}$. The third column reports the associated standard errors. The “Diff in Trend” column shows the point estimates for the difference in slope between the Large/High-Growth group and the other groups, while the last column reports the associated standard errors. ** $p < .05$.

B.4 Alternative specification for the common component

In this section, we show that our results about the trends in data processing hold under an alternative assumption about the common component in firm cash flows. Recall that in the baseline analysis, we imposed a single factor structure on the common component. Now, we allow for multiple aggregate factors with group-specific weights. Specifically, the correlated component is now given by

$$\tilde{\epsilon}_{fj1} = \sum_{h=1}^H \tilde{\beta}_{jh} \tilde{\epsilon}_{h1},$$

where H is the number of aggregate factors and $\tilde{\beta}_{jh}$ are the group-specific loadings.

Under these conditions, we can strip out the correlated components by taking out a group-year fixed effect from the observed cash flow and price variables. The rest of the estimation procedure to estimate PINF and its components remains unchanged. The results from this version are shown in Figure B2. Comparing it to our baseline results, reveals that the overall pattern of divergence emerges even under this alternative approach, indicating that our conclusions are not sensitive to how we adjust for common components.

C. Marginal Low-Growth of Information

C.1 Derivations

Interim expected utility, that is, after chosen information and prices are observed, is

$$\mathbb{E}[U_j^i | \mathcal{I}^i] = \frac{1}{2} \frac{(\mathbb{E}[V_{j1} - r P_{j1} | \mathcal{I}^i])^2}{\text{Var}[V_{j1} - r P_{j1} | \mathcal{I}^i]} = \frac{1}{2} \frac{(\Pi_j^i)^2}{\left(\frac{r}{r-g_j}\right)^2 (\Sigma_j^i)^{-1}} \tag{C.1}$$

Note that, from an ex ante perspective, Π_j^i is a random variable, since it is a function of the data observed by i . In our Gaussian setting, the posterior variance, Σ_j^i , depends only on second moments (which are known ex ante, that is, before data are observed). Ex ante expected utility therefore becomes

**Figure B2****Data divergence: Using group-year fixed effects**

This figure illustrates the structural estimation of Equation (13) after residualizing cash flows and prices using a group-year fixed effect. For each component, the lines plot a linear trend fitted to our structural estimates.

$$\mathbb{E}[U_j^i] = \mathbb{E}[\mathbb{E}[U_j^i | \mathcal{I}^i]] = \frac{1}{2} \frac{\mathbb{E}[(\Pi_j^i)^2]}{\left(\frac{r}{r-g_j}\right)^2} (\Sigma_j^i)^{-1} \quad (\text{C.2})$$

$$= \frac{1}{2} \left[\frac{(\mathbb{E}[\Pi_j^i])^2 + \text{Var}(\Pi_j^i)}{\left(\frac{r}{r-g_j}\right)^2} \right] (\Sigma_j^i)^{-1}, \quad (\text{C.3})$$

The unconditional mean and variance of expected profit per share can be computed directly from the equilibrium price function:

$$\mathbb{E}[\Pi_j^i] = \bar{\rho} \left(\frac{r}{r-g_j} \right)^2 \bar{\Sigma}_j \bar{x}_j. \quad (\text{C.4})$$

$$\text{Var}(\Pi_j^i) = B_j^2 \Sigma_{jp} + \left(\frac{r}{r-g_j} - B_j \right)^2 (\Sigma_{jd} - \Sigma_j^i) - 2 \left(\frac{r}{r-g_j} - B_j \right) B_j \Sigma_j^i \quad (\text{C.5})$$

The variance of expected profit depends, among other things, on the equilibrium pricing coefficient B_j and the noise in the price signal Σ_{jp} . Higher sensitivity to dividends or more noise leads to more ex ante variability in expected profits. Substituting the mean and variance of the expected profit per share into (C.3), we get

$$\begin{aligned}
 \mathbb{E}[U_j^i] &= \left[\bar{\rho}^2 \left(\frac{r}{r-g_j} \right)^4 \bar{\Sigma}_j^2 \bar{x}_j^2 \right] \frac{(\Sigma_j^i)^{-1}}{2 \left(\frac{r}{r-g_j} \right)^2} \\
 &+ \left[B_j^2 \Sigma_{jp} + \left(\frac{r}{r-g_j} - B_j \right)^2 (\Sigma_{jd} - \Sigma_j^i) - 2 \left(\frac{r}{r-g_j} - B_j \right) B_j \hat{\Sigma}_j^i \right] \frac{(\Sigma_j^i)^{-1}}{2 \left(\frac{r}{r-g_j} \right)^2} \\
 &= \left[\bar{\rho}^2 \left(\frac{r}{r-g_j} \right)^2 \bar{\Sigma}_j^2 \bar{x}_j^2 + \left(\frac{B_j}{\frac{r}{r-g_j}} \right)^2 \Sigma_{jp} + \left(1 - \frac{B_j}{\frac{r}{r-g_j}} \right)^2 \Sigma_{jd} \right] \frac{(\Sigma_j^i)^{-1}}{2} + H_j \\
 &= \left[\bar{\rho}^2 \left(\frac{r}{r-g_j} \right)^2 \bar{\Sigma}_j^2 \bar{x}_j^2 + \left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right)^2 \Sigma_{jp} + \left(\frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right)^2 \Sigma_{jd} \right] \frac{(\Sigma_j^i)^{-1}}{2} + H_j \\
 &= M_j \cdot (\Sigma_j^i)^{-1} + H_j
 \end{aligned}$$

where

$$M_j = \frac{1}{2} \left[\bar{\rho}^2 \left(\frac{r}{r-g_j} \right)^2 \bar{\Sigma}_j^2 \bar{x}_j^2 \right] + \frac{1}{2} \left[\left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right)^2 \Sigma_{jp} + \left(\frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right)^2 \Sigma_{jd} \right] \quad (\text{C.6})$$

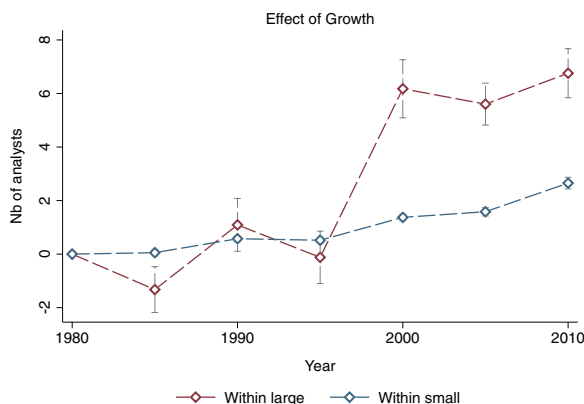
is the marginal value of information for asset j and the precision of the price signal and H_j is an equilibrium constant that does not depend on i 's information.

Note that M_j is a function, among other things, of the amount of data processed by the average investor (through $\bar{\Sigma}_j^2$ and Σ_{jp} terms). The value of information in (21) in the main text removes these effects by setting $\bar{\Sigma}_j^2 = \Sigma_{jd}$. The implications for Σ_{jp} comes from the pricing coefficients (see (11)). If no data are processed by others, then no information can be revealed in prices, so $B_j = 0$ and $\Sigma_{jp} = \infty$. At the same time, the term $\left(1 - \frac{\bar{\Sigma}_j}{\Sigma_{jd}} \right)^2$ becomes zero. Using L'Hospital's rule, we can show that the latter dominates and therefore, the product becomes zero in the no-information limit. Combining, the value of information M_j reduces to the expression for $V I_j$ in (21).

D. Evidence from Analyst Coverage

In this section, we present some evidence suggestive of increased data processing with respect to high-growth firms. We use the I/B/E/S database to estimate time trends in analyst coverage for different subsamples of firms. Formally, we regress the number of analysts at the firm-year level on a growth dummy, interacted with dummies for 5-year windows. We estimate this regression (allowing for year fixed effects) separately for large firms and for small firms. The coefficient for the growth dummy thus represents the relative coverage of high-growth firms. The results, presented in Figure D1 below, show a sharp increase in the relative coverage of high-growth firms. This is particularly striking for large firms and the timing of this increase lines up quite well with the results of our structural approach.

Of course, it is worth noting that analyst coverage is likely a rather crude measure of data precision. For one, the number of analysts doesn't capture variation in quality of data processing, both in the cross-section and over time. An analyst might be reporting mostly redundant or low-quality information that does little to reduce investor uncertainty (in fact, to the extent it disagrees with other analysts' forecasts, it might even seed uncertainty). Finally, analyst coverage also does not capture data processing done in-house by investors (e.g., hedge funds), which has arguably become more important over time. So while this evidence is reassuring and suggestive, it is not a substitute for a structural data precision measure.

**Figure D1****Analyst coverage has increased for high-growth firms**

The graph reports coefficients β_t from the following regression: $Number\ of\ analysts_{it} = \beta_t Growth_i \times Half-decade_t + \delta_t + e_{it}$, where $Growth_i$ is a dummy equal to one if the firm is a high-growth firm and $Half-decade_t$ is a dummy for each 5-year interval starting from 1985. We estimate the regression separately for large firms (the red line) and small firms (the blue line).

E. Price Informativeness: Additional Empirical Results

This appendix performs a number of exercises to show the evolution of price informativeness, defined as in Bai, Philippon, and Savov (2016). It is worth keeping in mind that these reduced-form patterns are difficult to interpret because they confound changes in information with variation in other characteristics, precisely why a structural approach is necessary. Having said that, these are still instructive and helps us connect our findings to various papers studying price informativeness.

Formally, we follow Bai, Philippon, and Savov (2016) and estimate the following specification:²¹

$$\frac{E_{f,j,t+s}^*}{A_{f,j,t}^*} = \alpha_j + \beta_{j,s} \cdot \ln\left(\frac{M_{f,j,t}^*}{A_{f,j,t}^*}\right) + \gamma_j \cdot X_{f,j,t} + \epsilon_{f,j,t+s} \quad (E.1)$$

and define price informativeness as

$$PINF_{j,s}^* = \beta_{j,s} \sigma_j^{M^*/A^*}, \quad (E.2)$$

where $\sigma_j^{M^*/A^*}$ denotes the (unconditional) standard deviation of $\ln\left(\frac{M_{f,j,t}^*}{A_{f,j,t}^*}\right)$. Finally, since we are interested in longer-term trends, we fit the following trendline (separately for each j):

$$PINF_{j,s,t}^* = \overline{PINF}_{j,s}^* \left(1 + Trend_{j,s} \cdot \frac{t - 1962}{2010 - 1962}\right) + e_{j,s,t} \quad (E.3)$$

The coefficient of interest is $\overline{PINF}_{j,s}^* \cdot Trend_{j,s}$, which describes how price informativeness changes over the period 1962–2010.

²¹ Throughout this appendix, we work with *unadjusted* prices and cash flows, that is, without taking out common components, in order to maintain comparability to Bai, Philippon, and Savov (2016) and the rest of the literature.

Table E1
The role of firm size in price informativeness

Dep. var	Price informativeness					
	S&P 500		Large firms		Small firms	
Sample (<i>j</i>)						
Horizon	<i>s</i> =3	<i>s</i> =5	<i>s</i> =3	<i>s</i> =5	<i>s</i> =3	<i>s</i> =5
	(1)	(2)	(3)	(4)	(5)	(6)
$PINF^*_{j,s} \cdot Trend_{j,s}$.016*** (.006)	.027*** (.006)	.0035 (.004)	.019*** (.0065)	-.052*** (.0038)	-.057*** (.0061)
$PINF^*_{j,s}$.033*** (.0023)	.038*** (.0036)	.041*** (.0023)	.048*** (.0038)	.043*** (.0018)	.054*** (.0029)
Observations	17,650	16,114	19,193	17,680	61,034	49,238
Sector FE	✓	✓	✓	✓	✓	✓
Firm controls	✓	✓	✓	✓	✓	✓

This table reports the results from estimating (E.3) for different subsamples of firms. Large firms are the 500 largest firms based on market capitalization. Small firms are the rest. Newey-West standard errors with four lags are in parentheses. ****p* <.01.

E.1 Price Informativeness for Largest (Smallest) Firms Has Been Rising (Falling)

To explore the connection between firm size and informativeness, we estimate $PINF^*_{j,s}$ and its trend for two subsamples: “largest” and “small” firms, where “largest” comprises the 500 largest firms, by market cap, and small the rest.

Table E1 reports the results for S&P 500 firms (columns 1 and 2), largest firms (columns 3 and 4) and small firms (column 5 and 6). The increase in price informativeness is very similar for S&P 500 firms and the set of largest firms, both for 3- (columns 1 and 3) and 5-year horizons (columns 2 and 4). By contrast, the price informativeness of small firms, which started from roughly the same levels as that of largest firms in 1962, fell sharply over this time period. These patterns are robust to alternative criterion for size: we also split the sample into deciles of size, and find that moving from the lowest decile to the highest decile of size implies a 17 fold increase in price informativeness (cf. Appendix E.3).

Next, we explore the relationship between growth and price informativeness. We classify firms based on their current book-to-market ratio, following Fama and French (1995). Specifically, firms in the bottom 30% by book-to-market are labeled “high-growth” firms and the top 30% “low-growth” firms. We then run our price informativeness regressions (E.1) separately for these two groups.

Columns 1 and 2 of Table E2 reveal that price informativeness declines for both high-growth and low-growth firms. However, when we split each category between large and small, we find that large high-growth firms show a significant increase (positive coefficient in column 4), while the small high-growth group displays the sharpest decline (column 3). In other words, high-growth firms drive both the rise in price informativeness for large firms and the declining trend for smaller firms. The informativeness for low-growth firms, both large and small, shows more modest declines. The rate of change in small low-growth firms’ (column 5) price informativeness is half that of small high-growth firms (column 3). The divergence is summarized in Figure E1, which plots the linear trends in price informativeness for large versus small firms (left panel) and for large/high-growth versus large-low-growth firms (right panel). Both panels exhibit divergence. Recall from Figure 3 that small firms, both high-growth and low-growth, show a declining trend.

E.2 Price Informativeness in the S&P 500

E.2.1 Price informativeness for all public firms (S&P 500 firms) has been falling (rising).

The two panels of Figure E2 plot the fitted values from (E.3) for the subsample of firms in the S&P 500 (left) and the universe of listed firms (right). The figures show that although informativeness rose for the S&P 500 firms, it fell for the market as a whole.

Table E2
The role of firm growth in price informativeness trends

Dep. var	Price informativeness ($s = 5$)					
	High-growth	Low-growth	High-growth– Small	High-growth– Large	Low-growth– Small	Low-growth– Large
Sample (j)	(1)	(2)	(3)	(4)	(5)	(6)
$PINF^*_{j,s} \cdot Trend_{j,s}$	-.035*** (.0083)	-.02*** (.0039)	-.058*** (.011)	.04*** (.01)	-.024*** (.0044)	-.01* (.0052)
$PINF^*_{j,s}$.052*** (.0052)	.014*** (.0024)	.054*** (.007)	.053*** (.0067)	.017*** (.0027)	.005* (.0029)
Observations	31,988	28,066	23,110	8,814	24,823	3,167
Sector FE	✓	✓	✓	✓	✓	✓
Firm controls	✓	✓	✓	✓	✓	✓

This table presents results from estimating Equation (E.3) for different subsamples of firms. Large refers to the 500 largest firms in our data; the rest are labeled Small. High-growth firms are those in the bottom 30% of the distribution of book-to-market; low-growth firms are in the top 30%. Newey-West standard errors, with four lags are in parentheses. *** $p < .01$.

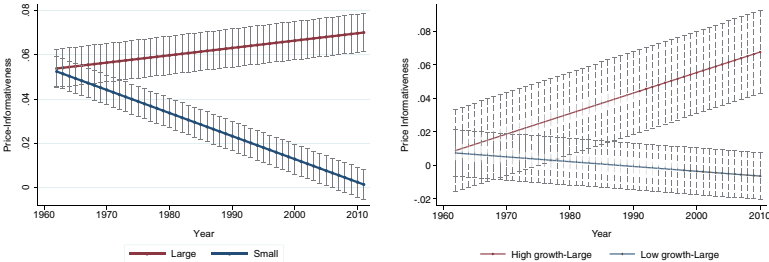


Figure E1
Large and small firms' price informativeness diverges

The plots show the trends in price informativeness for horizon $s=5$, estimated using (E.3), along with 95% confidence interval based on Newey-West standard errors with 4-year lags. Large refers to the 500 largest firms in our data; the rest are labeled Small. High-growth firms are those in the bottom 30% of the distribution of book-to-market; low-growth firms are in the top 30%.

Table E3 quantifies the magnitude of the divergent trends for S&P 500 and non-S&P 500 firms and shows that they are both statistically significant and economically large. $PINF^*_{j,s}$ reports the magnitude of the predictive power of stock prices for future cash flows at the beginning of our sample period. Because we normalize the time trend between zero and one, the coefficient for $PINF^*_{j,s} \cdot Trend_{j,s}$ can be directly interpreted as the total evolution of price informativeness over the period. For the S&P 500 sample, price informativeness at the 5-year horizon rose by 70% (0.026/0.038). For the non-S&P 500 firms, it fell by around 80%. In all cases, the evolution is significant at the 1% level.

To explore whether there is something specific to firms in the S&P 500, we perform two different tests. First, we look at firms that have never been included in the S&P 500 but are relatively close in terms of market capitalization and size. These firms exhibit a rise in price informativeness nearly identical to that of the S&P 500 firms (though the levels of price informativeness are somewhat different). This suggests that the rising trend in price informativeness has more to do with firm characteristics (like size) rather than inclusion in the S&P 500 *per se* (though being part of the index does increase the *level* of informativeness somewhat).

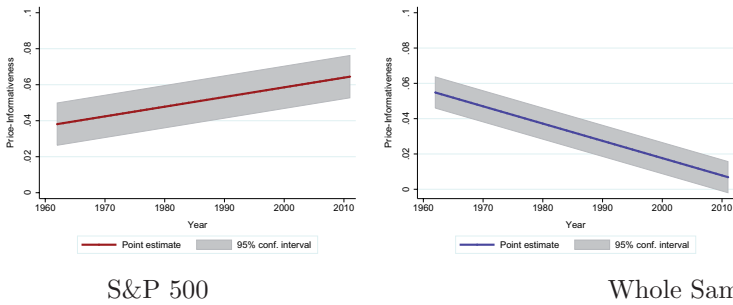


Figure E2

Price informativeness is falling (rising) for all public firms (S&P 500 firms)

The plots depict the trends in price informativeness, estimated using (E.3), along with 95% confidence interval based on Newey-West standard errors with five lags. The left panel depicts S&P 500 nonfinancial firms, while the right shows the results for the whole sample.

Table E3
Price informativeness grew (fell) for S&P 500 (other) firms

	1960s	1970s	1980s	1990s	2000s	2010s
Number of firms						
Small high-growth	1,699	4,739	7,224	9,253	6,444	3,505
Large high-growth	1,696	4,229	6,270	7,963	5,662	3,327
Small low-growth	1,734	4,664	7,229	9,153	6,382	3,472
Large low-growth	1,653	4,040	6,146	7,742	5,534	3,272
Average assets (\$ millions)						
Small high-growth	125	173	109	175	410	599
Large high-growth	2,697	3,510	3,521	8,661	12,928	13,802
Small low-growth	517	565	852	2,140	4,478	5,398
Large low-growth	6,129	11,592	15,726	22,003	52,550	61,588

This table shows the estimates of (E.3) for different subsamples of firms. Newey-West standard errors, with four lags are in parentheses. *** $p < .01$.

We also looked at firms that were in the S&P 500 only for a part of our sample period. We estimate two separate specifications of Equation (E.1): one for the period of the firm life when it is in the S&P 500 and one for when it is not. We find that, among the sample of firms that are in the S&P 500 at some point in their life, the trend in price informativeness is similar for firms currently in and out of the S&P 500. In levels, price informativeness is actually higher when a firm is not in the S&P 500, than when they are in.

E.3 Price Informativeness by Size

In this subsection, we first document evolution of firm size during our sample period, and then show that price informativeness varies systematically by size.

Figure E3 show that S&P 500 firms got larger, relative to non-S&P 500 firms. Here, we use market capitalization as our measure, but the pattern looks similar with assets as well. As we showed in Section 4 in the main text, size is a key determinant of the value of information, so this diverging trend in size helps explain the diverging trends in data.

To study the variation of price informativeness by size, we pool all firm-year observations and construct deciles of firm size (defined as market value in 2009 dollars). We then run the cross-sectional regression (E.1) within each bin; that is, the subscript j now refers to a size bin and estimate $PINF_{j,s}^*$. The results, presented in Figure E4, show a clear pattern: the informativeness

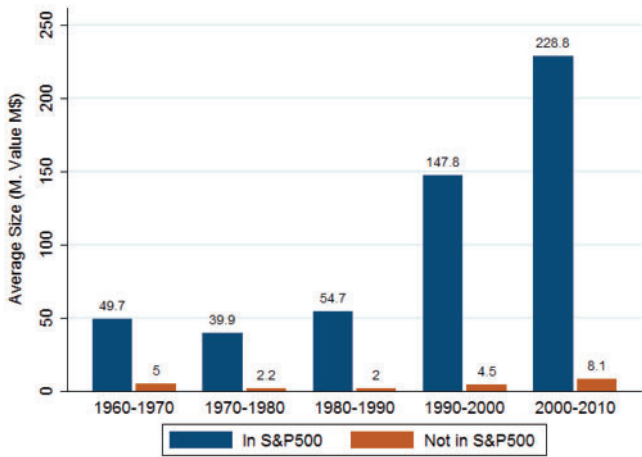


Figure E3
S&P 500 firms became larger relative to non-S&P 500 firms
The graph shows the average size of S&P 500 and non-S&P 500 firms over time. Size is defined as firms' total market value in 2009 dollars. The sample contains publicly listed nonfinancial firms from 1960 to 2010.

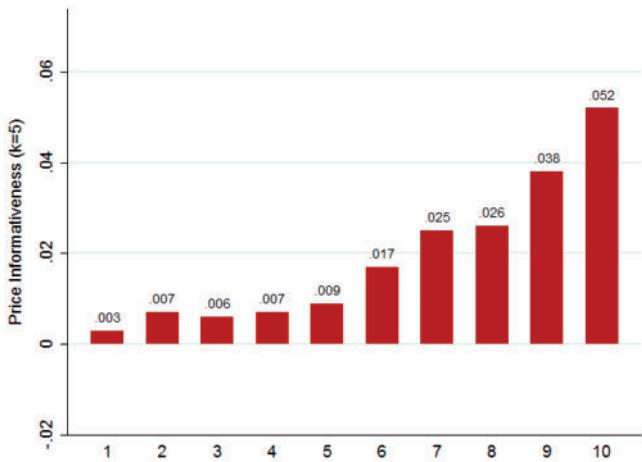


Figure E4
Price informativeness by decile
The figure shows the average $PINF_{j,s,t}^*$, defined as in (E.2), over the entire sample for each size decile. We run the regression in (E.1) for each year $t = 1962, \dots, 2010$ with horizon $s = 5$ for each size decile. The sample contains publicly listed nonfinancial firms from 1962 to 2010.

of large firms is significantly higher than those of smaller firms, especially for those at the very top.

E.4 Other Possible Data Groupings

One potential concern with our analysis is that growth and size are not the characteristics that are driving these trends, but are correlated with other, more relevant firm characteristics. In this subsection, we discuss a couple of other groupings of firms that might help dig into this further.

E.4.1 Technology firms. A potential explanation for the decrease in informativeness for the market as a whole is that the share of firms, whose shares are more difficult to price—specifically high-tech firms—has increased over time. Could the increased prevalence of technology firms also explain divergence? However, we find that quantitatively, the rise of such firms explains little of the divergence in price informativeness, because the technology-related time trends in the large firm and small firm samples were not sufficiently different.

We use R&D intensity (R&D spending scaled by assets) as a proxy for high tech intensity. First, we sort the full sample of firm-year observations into deciles of R&D intensity. We then estimate price informativeness for each decile, using the same method as before. We find that price informativeness declines strongly with R&D intensity, as we conjectured.

Next, we analyze changes in R&D composition in the cross-section. We use inclusion in the S&P 500 as our indicator of being a large firm. In both the S&P 500 and the non-S&P 500 sample, the fraction of firms investing more in R&D has increased steadily. The share of high-tech firms has grown slightly more rapidly in the full sample than in the S&P 500 sample. Until the early 1980s, the high-tech shares for S&P 500 and non-S&P 500 firms track each other closely. Some signs of divergence creep up in the mid-1980s, when the share of high-tech firms increases more in the whole sample, essentially driven by a rapid entry rate of tech firms. But then, in the early 2000s, the share of tech firms in the S&P 500 increases and converges to that of the non-S&P 500 sample. Thus, a clear trend does not surface in the tech composition of the different subsamples. We therefore conclude that prevalence of tech firms, while it may explain the average decline in informativeness, cannot explain the cross-sectional divergence.

Note also that our structural approach explicitly adjusts the effect of differences in fundamentals, for example, a more volatile or faster-growing cash flow. So, to the extent that technology firms are different for these reasons, our analysis in that section adjusts for technology intensity and finds divergence.

E.4.2 Market power. Recent work suggests that market power is rising in the U.S. economy over the last few decades. In Kacperczyk, Nosal, and Sundaresan (2018), market power considerations reduces price informativeness: large investors with price impact trade less aggressively on their information, leading to lower price informativeness. This could be a potential explanation for the overall decline in price informativeness. This would imply that price informativeness we estimate is a lower bound (as is our structural measure of data). But, for this to explain why only large, high-growth firms have much more informative prices than they used to, we would have to argue that the market for those stocks has become much more competitive over time. To the best of our knowledge, there is no evidence that suggests enormous increases in competition in some equity markets and the evaporation of competition in others.

References

- Abis, S. 2017. Market timing in Bayesian portfolio optimization. Working Paper, Columbia GSB.
- . 2018. Man vs. machine: Quantitative and discretionary equity management. Working Paper, Columbia University.
- Admati, A. 1985. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53:629–57.
- Andries, M., and V. Haddad. 2020. Information aversion. *Journal of Political Economy* 125:1901–39.

- Bai, J., T. Philippon, and A. Savov. 2016. Have financial markets become more informative? *Journal of Financial Economics* 122:625–54.
- Ben-Rephael, A., B. Carlin, Z. Da, and R. Israelsen. 2021. Information consumption and asset pricing. *Journal of Finance* 76:357–94.
- Berk, J., and R. C. Green. 2004. Mutual fund flows and performance in rational markets. *Journal of Political Economy* 112:1269–95.
- Bond, P., and H. Eraslan. 2010. Information-based trade. *Journal of Economic Theory* 145:1675–703.
- Bond, P., A. Edmans, and I. Goldstein. 2012. The real effects of financial markets. *Annual Review of Financial Economics* 4:339–60.
- Brogaard, J., H. Nguyen, T. Putnins, and E. Wu. 2018. What moves stock prices? The role of news, noise and information. Working Paper, University of Washington.
- Chemmanur, T. J., and A. Yan. 2019. Advertising, attention, and stock returns. *Quarterly Journal of Finance* 9:10.1142/S2010139219500095.
- Chordia, T., C. Green, and B. Kottimukkalur. 2018. Rent seeking by low latency traders: Evidence from trading on macroeconomic announcements. *Review of Finance Studies* 31:4650–87.
- Crouzet, N., I. Dew-Becker, and C. Nathanson. 2020. On the effects of restricting short-term investment. *Review of Financial Studies* 33:1–43.
- David, J., H. Hopenhayn, and V. Venkateswaran. 2016. Information frictions, misallocation and aggregate productivity. *Quarterly Journal of Economics* 131:943–1005.
- Dávila, E., and C. Parlatore. (2019). Volatility and informativeness. Working Paper, Yale University.
- Davis, S. J., and J. Haltiwanger. 2019. Dynamism diminished: The role of credit conditions. Working Paper, University of Chicago.
- Dessaint, O., T. Foucault, L. Fresard, and A. Matray. 2019. Noisy stock prices and corporate investment. *Review of Financial Studies* 32:2625–72.
- Dow, J., I. Goldstein, and A. Guembel. 2017. Incentives for information production in markets where prices affect real investment. *Journal of the European Economic Association* 15:877–909.
- Dugast, J., and T. Foucault. 2018. Data abundance and asset price informativeness. *Journal of Financial Economics* 130:367–91.
- Durnev, A., R. Morck, and B. Yeung. 2004. Value-enhancing capital budgeting and firm-specific stock return variation. *Journal of Finance* 59:65–105.
- Edmans, A., S. Jayaraman, and J. Schneemeier. 2017. The source of information in prices and investment-price sensitivity. *Journal of Financial Economics* 126:74–96.
- Fama, E. F., and K. French. 1995. Size and book-to-market factors in earnings and returns. *Journal of Finance* 50:131–55.
- Farboodi, M., and L. Veldkamp. 2020. Long-run growth of nancial data technology. *American Economic Review* 110:2485–523.
- Glode, V., R. Green, and R. Lowery. 2012. Financial expertise as an arms race. *Journal of Finance* 67:1723–59.
- Goldstein, I., E. Ozdenoren, and K. Yuan. 2013. Trading frenzies and their impact on real investment. *Journal of Financial Economics* 109:566–82.
- Gschwandtner, A. 2012. Evolution of prot persistence in the usa: Evidence from three periods. *Manchester School* 80:172–209.
- Guo, Y., and L. Mota. Forthcoming. Should information be sold separately? Evidence from mifid ii. *Journal of Financial Economics*.

- Hong, H., and M. Kacperczyk. 2010. Competition and bias. *Quarterly Journal of Economics* 125:1683–725.
- Hong, H., T. Lim, and J. C. Stein. 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance* 55:265–95.
- Kacperczyk, M., J. Nosal, and L. Stevens. 2019. Investor sophistication and capital income inequality. *Journal of Monetary Economics* 107:18–31.
- Kacperczyk, M., J. Nosal, and S. Sundaresan. 2018. Market power and informational efficiency. Working Paper, Imperial College London.
- Katz, M., H. Lustig, and L. Nielsen. 2017. Are stocks real assets? Sticky discount rates in stock markets. *Review of Financial Studies* 30:539–87.
- Kelly, B., and A. Ljungqvist. 2012. Testing asymmetric-information asset pricing models. *Review of Financial Studies* 25:1366–413.
- Davila, E., and C. Parlato. 2016b. Identifying price informativeness. Working Paper, NYU.
- Kyle, A. S., and J. Lee. 2017. Toward a fully continuous exchange. *Oxford Review of Economic Policy* 33:650–75.
- Maćkowiak, B., and M. Wiederholt. 2009. Optimal sticky prices under rational inattention. *American Economic Review* 99:769–803.
- Martineau, C. 2017. The evolution of market price efficiency around earnings news. Working Paper, University of Toronto.
- Mondria, J., T. Wu, and Y. Zhang. 2010. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics* 82:85–95.
- Ozdenoren, E., and K. Yuan. 2008. Feedback effects and asset prices. *Journal of Finance* 63:1939–75.
- Pástor, L., and R. F. Stambaugh. 2012. On the size of the active management industry. *Journal of Political Economy* 120:740–81.
- Ranco, G., D. Aleksovski, G. Caldarelli, M. Grcar, and I. Mozetic. 2015. The effects of twitter sentiment on stock price returns. *PLoS ONE* 10:10.1371/journal.pone.0138441.
- Sims, C. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50:665–90.
- Stambaugh, R. 2014. Presidential address: Investment noise and trends. *Journal of Finance* 69:1415–53.
- Van Nieuwerburgh, S., and L. Veldkamp. 2009. Information immobility and the home bias puzzle. *Journal of Finance* 64:1187–215.