Valuing Financial Data

Maryam Farboodi[*] Dhruv Singal[†]

Laura Veldkamp[‡] Venky Venkateswaran[§]

April 16, 2024[¶]

Abstract

How should an investor value financial data? The answer is complicated because it depends on the characteristics of all investors. We develop a sufficient statistics approach that uses equilibrium asset return moments to summarize all relevant information about others' characteristics. It can value data that is public or private, about one or many assets, relevant for dividends or for sentiment. While different data types, of course, have different valuations, heterogeneous investors also value the same data very differently, which suggests a low price elasticity for data demand. Heterogeneous investors' data valuations are also affected very differentially by market illiquidity.

[*]MIT Sloan, NBER and CEPR; farboodi@mit.edu

[†]Columbia Business School; dhruv.singal@columbia.edu

[‡]Columbia Business School, NBER, and CEPR; lv2405@columbia.edu

[§]NYU Stern School of Business and NBER; vvenkate@stern.nyu.edu

[¶]We thanks our editor, Itay Goldstein, as well as the two anonymous referees for a very constructive set of comments. We are also grateful to Adrien Matray, Vincent Glode, Cecilia Parlatore and Eduardo Davila for valuable conversations and suggestions. Keywords: Data valuation, portfolio theory, information choice.

Investment management firms are gradually transforming themselves from users of small data and simple asset pricing models to users of big data and computer-generated statistical models. Amidst this transformation, investors' strategic focus is shifting from the choice of pricing model to the choice of data they acquire. A key question for modern financial firms is: How much should they be willing to pay for a stream of financial data? This paper devises and puts to use a methodology to estimate this dollar value, based the investor's own characteristics, without needing to know the characteristics of others.

From information-based theories, we know many qualitative features of firms that make data valuable – large stocks, growth stocks, stocks with risky payoffs, assets that are sensitive to news, assets that others are uninformed about. After all, data is simply a stream of digitized information. But investors differ. An investor with a large portfolio values data more, while an investor who invests in a restricted set of assets values data less. Investors with different investment styles value data differently. An investor with lots of other data is less willing to pay for additional data, while an investor who trades more frequently might value data more or less. The magnitudes of all these effects depend on the asset market equilibrium, which in turn depends on the characteristics of every other investor. To make matters more complex, we also know that illiquidity or price impact of a trade make information less valuable (Kacperczyk, Nosal, and Sundaresan, 2021), but how this interacts with investor heterogeneity, quantitatively, is less understood.

Our simple procedure to estimate the value of any data series, to an investor with specific characteristics, reveals enormous dispersion in how different investors value the same data. The dispersion in private valuations matters for our understanding of data markets because when values are highly dispersed and the market price of data changes, few customers have valuations between the old and new price. So, few customers change their data demand in response to the price change. This is a low price elasticity of aggregate data demand.

It is important to point out that our procedure leads to an estimate of private *value* to an investor, which could be different from a transaction price that one might observe when

data is sold. Knowing the private values of market participants allows one to trace out a demand curve for data. Some investors would have values greater than the equilibrium price, some less. This is like a shopper determining how much they value a sweater. Knowing that the sweater's market price is $50 does not make that the shopper's value of the sweater – it might be the wrong color or size. Alternatively, the shopper might be willing to pay $100 for the sweater and still not buy it because they find a similar sweater for less. Understanding how customers (investors) value a product (a data set) is different from calculating a market clearing or equilibrium price. Valuations are important because they allow us to evaluate consumer surplus and welfare, teach us about demand elasticity, markups and market competition, and allow one to ask if observed transactions prices are efficient.

Our measurement approach relies on sufficient statistics which are easily computable. While our measure is based on a model, we do not need to estimate most model parameters to arrive at a data value. In Section 1, we set up a noisy rational expectations model with rich heterogeneity in investors, assets and data types and derive the expected utility of data in dollar amounts. These investors use data to seek alpha. They control for risk factors, forecast which assets have excess returns and invest in those assets. We show that a few sufficient statistics—average conditional and unconditional returns, variances and forecast errors—are all that is needed to value a stream of data.

The fact that these return-based statistics are sufficient is surprising. On the surface, they seem to be missing the well-known result that private information is more valuable than public information. Furthermore, they seem to miss the extent to which an investor's data is partially revealed to others through the price level. However, our statistics do account for both forces and the results show why. Public information and the part of information leaked through prices affect prices, but do not forecast returns. Looking at the reduction in the conditional variance of returns correctly values data for its information content, over and above what can be learned from prices, and accounting for what others will learn from prices about our data as well. Our sufficient statistics are also a valid measure, regardless of

how heterogeneous other investors' preferences, data or investment styles are. They can be used to value data about asset fundamentals or about sentiment. Finally and importantly, a version of these statistics can be used in imperfectly competitive markets as well.

One could apply this tool to any finance-relevant data series, or any bundle of data series – all it requires is knowledge of the purchasing investor's characteristics and access to a history of market prices and data realizations. We present a small number of examples that highlight the importance of accounting for investor heterogeneity in data valuation.

In Section 3, we compute the dollar value of median analyst forecasts of earnings growth for investors with different wealth levels, different investment styles and facing different market conditions. These exercises highlight the flexibility of our approach and its ability to accommodate various dimensions of heterogeneity. Our first exercise explores the role of investor wealth and risk preferences. To do this, we consider two investors with the same relative risk aversion and different initial wealth levels. This implies that the wealthier investor has lower absolute risk aversion and as a result, values the same data by more. But, the extent depends on market conditions, i.e. on whether their trades have price impact or not. When markets are competitive, i.e. a trade has no impact on the market price, data values increase sharply with wealth: an investor with $250 million in initial wealth values data by almost 300 times compared to one with $0.5 million. Accounting for price impact, in line with empirical estimates, dramatically reduces the value of data for all investors, but has noticeably larger effects on the investor with higher wealth/lower absolute risk aversion. This illustrates a general pattern we see – there is enormous heterogeneity in willingness to pay for data, that is substantially tempered by a modest degree of market illiquidity.

The high sensitivity of data to price impact is interesting in its own right. It suggests that market liquidity matters greatly for the value of financial data. Small changes in market conditions can thus lead to large variation in data value and through that, in the valuation of firms whose main asset is financial data. This suggests a new avenue of how liquidity effects in asset markets. We typically think of market liquidity as something that affects

only the value of financial assets, not directly affecting the real value of a firm. As data becomes a more important asset for financial firms, the prices of financial firms may become increasingly sensitive to market liquidity.

Incorporating price impact also uncovers a novel insight: Inelastic asset demand can lead to more elastic data demand. This is because price impact causes investors to reduce the sensitivity of their trading decisions to prices, implying a less elastic asset demand. Price impact also makes data valuations less heterogeneous by lowering data value most significantly for investors with the highest data valuations to begin with, flattening the data demand curve and contributing to high price elasticity.

Our second exercise considers investors with different investment styles. Specifically, we analyze the value of analyst forecast data for investors who trade only in a single portfolio, such as the S&P 500 portfolio, or a portfolio consisting of only small stocks, only large stocks, only growth stocks or only value stocks. Our benchmark is an investor who trades all five of these portfolios. Because each of these types uses a piece of data differently, they value the same piece of information differently. We find that investors in large and growth stocks (as well as investors who trade all five portfolios) value analyst forecast data substantially more than a value or small-stock investor.

Our third exercise quantifies how much the value of analyst forecast data depends on what other data is in an investor's database. We find considerable variation in data values when we vary the other data variables used. In general, the more series we add to the investor's information set, the lower is the value gained by having access to analyst forecasts and these effects are sizable. This result illustrates the importance of accounting for many facets of investor heterogeneity. It also suggests that this dimension of heterogeneity can induce sizable variation in data valuations.

Our fourth exercise considers the effect of trading horizon on the value of analyst forecast data. Our toolkit can easily accommodate such differences with higher frequency observations on the data series and asset returns. We illustrate this by computing the value of data

to an investor who trades over an annual horizon (our baseline calculations are for a quarterly horizon). We find that a longer horizon makes analyst forecast data somewhat more valuable, i.e., the data turn out to be more useful in forecasting returns over shorter horizons. It is worth highlighting that this exercise is about trading horizon, not frequency—it is possible that an investor who trades or rebalances his portfolio more frequently ascribes a higher value to the data compared to one who trades less often. In principle, our procedure can be extended to this type of heterogeneity as well, but in part due to data limiations, we do not explore it in this paper.

In Section 4, we explore how investors with different characteristics value macroeconomic information. In contrast to the analyst forecast data, we find less variation across different investment styles for this type of data. The estimated dollar values are sizeable, suggesting all investors in general find macroeconomic information quite useful for portfolio choice. Again, market illiquidity not only decreases value of data to all investors, but also significantly compresses the heterogeneity in data valuations.

As such, these exercises not only highlight how our toolkit can be applied in practice but also yield new insights about financial asset and data markets.

Relationship to the Literature. Data is information. Therefore, our approach to valuing financial data draws primarily on the literature exploring information in financial markets. A few papers have examined the value of information or skill, for a representative agent or in an economy with one aggregate risk (Kadan and Manela, 2019; Savov, 2014; Dow, Goldstein, and Guembel, 2017; Morris and Shin, 2002). Kacperczyk, Nosal, and Sundaresan (2021), Kyle and Lee (2017), and Kyle (1989) add imperfect competition. What we add is a richer asset structure, a richer information structure, but most importantly, heterogeneous investors who value information differently. This last ingredient is essential to understand what the aggregate data demand function looks like. The study of data demand complements work examining the different ways in which data is supplied (Admati and Pfleiderer, 1986, 1987).

Enriching the information structure to allow for public, private or correlated signals is

also important for real-world measurement. Such rich information structures are commonly studied in settings with quadratic payoffs (Ozdenoren and Yuan, 2008; Albagli, Hellwig, and Tsyvinski, 2014; Amador and Weill, 2010). But they have substantially complicated previous asset market models to the point that most authors assume fully private (Barlevy and Veronesi, 2000; He, 2009; Kondor, 2012) or fully common (Grossman and Stiglitz, 1980) information.[1] In addition, investors may choose between asset valuation-relevant data or data about other investors' order flow (Farboodi and Veldkamp, 2017). The idea that all these types of information can be valued with one set of sufficient conditions is a new idea that substantially broadens the empirical applicability of these tools.

The main point of the paper is to show that heterogeneity in investor characteristics matters for data valuations. Some version of all these characteristics exist in some noisy rational expectations model (Kacperczyk, Nosal, and Sundaresan, 2021; Peress, 2004; Mondria, 2010), most of which look daunting to estimate.[2] This project shows that, despite all these degrees of heterogeneity among investors, data types and equilibrium effects, there is a simple procedure to compute a value for data.

Campbell and Thompson (2008) propose a data valuation procedure that, like ours, also makes use of conditional means and the variances of returns. Our work advances this conversation, by accounting for the revelation of information through prices, showing how to incorporate many, correlated assets and signals, exploring public and private data, accounting for market elasticity or price impact, and allowing for wealth effects that are consistent with common utility specifications. Our results discuss the quantitative significance of each of these innovations.

Measures of the information content of prices, like those in Bai, Philippon, and Savov (2016) and Davila and Parlatore (2021) are used to infer how much the average investor in an asset knows. Such measures are related, in that they arise from a similar noisy rational

[1] Exceptions include Goldstein, Ozdenoren, and Yuan (2013) and Sockin (2015).

[2] Heterogeneity also arises in micro models like (Bergemann, Bonatti, and Smolin, 2016), who value information in a bilateral trade, where sellers do not know buyers' willingness to pay, but without the equilibrium considerations about what others know.

expectations framework. But they answer a question about the quantity of information, not its value. Farboodi, Matray, Veldkamp, and Venkateswaran (2019)'s "initial value" of a unit of precision is not the value an investor would pay and is not heterogeneous. Our sufficient statistics approach is more relevant for demand estimation, much simpler to estimate and more robust to heterogeneity.

1 A Framework for Valuing Data

Since data is information, we build on the standard workhorse model of information in financial markets, the noisy rational expectations framework, in which investors use signals and the information in asset prices to select high excess return portfolios, controlling for risk. To the framework, we add long-lived assets, imperfect competition, wealth effects, investment styles, public, private or partly public signals and arbitrary correlation between assets and between various signals. We include these features because each one affects the value of information. Model extensions consider data about sentiment or order flow.

Our contribution is not a new *model per se*, but rather to show how to estimate data valuations using a rich and flexible theoretical framework. Of course, this is not the simplest model to arrive at a sharp result on data valuations. If anything, it is the opposite: our goal is to show how, despite all the richness and heterogeneity, the value of data can still be reduced to a few sufficient statistics that are easy to compute. Later, we justify this rich modeling of investor heterogeneity by showing that it has a significant impact on data valuations.

Assets There are N distinct risky assets in the economy indexed by j, with net supply given by $\bar{x}$. Each of these assets are claims to stream of dividends $\{d_{jt}\}_{t=0}^{\infty}$, where the vector d_t is assumed to follow the auto-regressive process

$$d_{t+1} = \mu + G(d_t - \mu) + y_{t+1}.$$

Here, the exogenous dividend innovation shock $y_{t+1} \sim \mathcal{N}(0, \Sigma_d)$ is assumed to be i.i.d. across time.[3] We use subscript t for variables that are known before the end of period t. Thus, the dividend d_{t+1} and its innovation shock y_{t+1} both pertain to assets that are purchased in period t; both these shocks are observed at the end of period t.

Investors and Investment Styles In each period t, n overlapping generations investors, $i \in [0, 1]$, are born, observe data, and make portfolio choices. The number of investors may be finite, which implies that markets are imperfectly competitive. We will also consider the limiting economy as n becomes infinite. In the following period $t + 1$, investors sell their assets, consume the dividends and the proceeds of their asset sale and exit the model. Each investor i born at date t has initial endowment $\bar{w}_{it}$ and utility over total, end-of-life consumption c_{it+1}.

Many investors describe their strategy as small-firm investing or value investing, which limits the assets they hold. In order to account for the role of investment strategy in data value, we allow an investor i to be subject to an investment style constraint, which limits the set of risky assets they can purchase. For each investor i, we denote the set of all portfolios over investable assets as $\mathcal{Q}_i$. The matrix θ_i is an $m_i \times N$ matrix of zeros and ones, where $m_i \equiv |\mathcal{Q}_i|$ is the number of investable assets for investor i. Each row of θ_i has a single 1 entry, with all other entries zero. Assume an arbitrary order on the N risky assets, if the asset indexed j in the entire set of N assets is the k-th asset in investor i's style class, then that asset is investable and the k-th row of θ_i will have jth column entry equal to 1.[4]

At date t, investors choose their portfolio of risky assets, which is a vector $q_{it} \in \mathcal{Q}_i$ of the number of shares held of each asset. They also choose holdings of one riskless asset with

[3]Normal payoffs and information, while standard in this literature, are not very realistic. Appendix D shows how to approximate a solution to this model in a case where payoffs and signals are not normal, but are skewed.

[4]Following, Koijen and Yogo (2019), we do not model the source of the constraint. Our formulation implies that we consider sets $\mathcal{Q}_i$ that either set the holdings of some assets to zero, or allow the entire real line. For example, long-only portfolios would restrict $\mathcal{Q}_i$ to the non-negative realm of Re^N. Of course, it is possible that an investor is unrestricted. If so, $\mathcal{Q}_i = \mathrm{Re}^N$.

return r, subject to budget constraint. Investor maximization can be written as

$$\max_{q_{it}} \mathbb{E}\left[U(c_{it+1})\right]$$
$$c_{it+1} = r\left(\overline{w}_{it} - q'_{it}\theta_i p_t\right) + q'_{it}\theta_i\left(p_{t+1} + d_{t+1}\right). \tag{1}$$

Data According to Bayes' law, the information investors learn from prices, as well as their private data, public data and any correlated information can be linearly combined into one composite signal. Specifically, this combination of private, public and price information is equivalent to getting an unbiased signal s_{it} about the dividend innovation y_{t+1}, with private signal noise ξ_{it} and public signal noise z_{t+1}.

$$s_{it} = y_{t+1} + \zeta_{it} z_{t+1} + \xi_{it}$$

The term $z_{t+1} \sim \mathcal{N}(0, \Sigma_z)$ comes from the noise in public component of the any data. It is iid across time, with precision Σ_z^{-1}. This public signal noise z_{t+1} pertains to assets that are purchased in period t and is observed at the *end* of period t. If investor i learned nothing from any prices or public sources of information at date t, then $\zeta_{it} = 0$ and this becomes a standard private signal. Similarly, $\xi_{it} \sim \mathcal{N}(0, K_{it}^{-1})$ is the noise in the private component of the signal (iid across individuals and time), which has the precision K_{it}, orthogonal to the noise of the public component.[5]

External Demand Some source of noise in prices is necessary to explain why some investors know information that others do not. Noise could come from hedging motives, estimation error, cognition errors or sentiment. Noise traders buy x_{t+1} shares of the asset, where $x_{t+1} \sim N(0, \Sigma_x)$ is independent of other shocks in the model and independent over

[5]This is equivalent to a setting where investors learn from prices. We recognize that the precision K will therefore depend on the equilibrium price coefficients. Previous versions of the paper also spelled out the equivalence between this form and a setting where each investor has access to H distinct data sources. Signals from each of these data sources (indexed by h) provides information about dividend innovations y_{t+1}, from a linear combination ψ_h of assets: $\eta_{iht} = \psi_h y_{t+1} + \Gamma_h e_{it}$.

time. The noise can be arbitrarily small, as long as $\Sigma_x > 0$.

Equilibrium An equilibrium is a sequence of prices $\{p_t\}_{t=0}^{\infty}$ and portfolio choices $\{q_{it}\}_{t=0}^{\infty}$, such that

1. At the beginning of each period t, all investors have the information set $\mathcal{I}_t^- = \{\mathcal{I}_{t-1}, y_t, d_t, x_t, z_t\}$, where $\mathcal{I}_{t-1}$ is the information set of the average investor at time $t-1$. The dividend d_{t+1} its innovation shock y_{t+1} and the noise trader demand x_{t+1} are observed at the end of period t and are included in $\mathcal{I}_{t+1}^-$.

2. Investors use Bayes' Law to combine prior information $\mathcal{I}_t^-$ with data and price information p_t to update beliefs. The information set at the time of portfolio choice is equivalent to $\mathcal{I}_{it} = \{\mathcal{I}_t^-, s_{it}, p_t\}$.

3. Investors choose their risky asset investment $q_{it} \in \mathcal{Q}_i$ to maximize $\mathbb{E}\left[U(c_{it+1})|\mathcal{I}_{it}\right]$, taking the actions of other investors as given, subject to the budget constraint (1).

4. At each date t, the price vector p equates demand plus noise x_{t+1} to $\bar{x}$ units of supply:

$$\sum_i q_{it} + x_{t+1} = \bar{x} \qquad \forall t. \tag{2}$$

Equilibrium Solution To solve the model, we assume that investors have mean-variance preferences over their end-of-period wealth.[6] This allows us to write the conditional expected utility at time t as

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i \mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2}\mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right]. \tag{3}$$

Here, ρ_i denotes absolute risk aversion for investor i.

[6]In Appendix C, we use the small shock approximation introduced in Peress (2004) to interpret mean-variance utility as a second-order approximation to a more general class of utility functions.

For a perfectly competitive market ($n \to \infty$), the equilibrium price schedule is linear in current dividend d_t, future dividend innovations y_{t+1}, demand shocks x_{t+1} and the noise in public data z_{t+1} (see Appendix A):

$$p_t = A_t + B(d_t - \mu) + C_t y_{t+1} + D_t x_{t+1} + F_t z_{t+1}. \tag{4}$$

Note that while our setting is dynamic (the assets are long-lived), the assumption of 2-period-lived investors leads to a simple Markov structure for the equilibrium price. The equilibrium price only depends on the current innovations, i.e., there are no dynamic hedging motives. This is motivated by the standard approach in the NREE literature, Veldkamp (2011), to keep the model tractable.

Mapping Data Utility to Sufficient Statistics Our first result derives the unconditional expectation of (3), in terms of means and variances of the vector of asset profits. Those profits, Π_{it}, for investor i's feasible investment set, are defined as

$$\Pi_{it} \coloneqq \theta_i \left[p_{t+1} + d_{t+1} - r p_t \right]. \tag{5}$$

Then, we express utility as an indirect expected utility function $\tilde{U}$ which takes an information set $\mathcal{I}_{it}$ (data) as its argument.

Lemma 1. *In a competitive market ($n \to \infty$), investor expected utility can be expressed as*

$$\tilde{U}(\mathcal{I}_{it}) = \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} \mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2} Tr\left[\mathbb{V}\left[\Pi_{it}\right] \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\rho_i \bar{w}_{it} \tag{6}$$

where Tr is the matrix trace and $\bar{w}_{it}$ is investor i's exogenous endowment.

Proof is in Appendix B. Equation (6) illustrates the basis for our measurement strategy. We will estimate $\tilde{U}(\mathcal{I}_{it})$ with and without the piece of data to be valued and take a difference.

The first term is the expected profit on individual i's portfolio. The role of more or better

data is to reduce conditional variance $\mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]$. In other words, an investor's utility rises with data if she can use the data to make forecasts with smaller squared forecast errors. Smaller forecast errors are valuable because they allow the investor to buy more of assets that will ultimately have higher returns

In an imperfectly competitive market, expected utility uses price-impact-adjusted variances, as proven in Appendix B.

Lemma 2. *Unconditional expected utility, for an investor with price impact dp/dq_i is*

$$\tilde{U}(\mathcal{I}_{it}) = \mathbb{E}[\Pi_{it}]' \hat{V}_i^{-1} \mathbb{E}[\Pi_{it}] + Tr\left[\left(\mathbb{V}[\Pi_{it}] - \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]\right) \hat{V}_i^{-1}\right] + r\rho_i \bar{w}_{it}. \tag{7}$$

where $\hat{V}_i^{-1} := \tilde{V}_i^{-1}\left(1 - \frac{1}{2}\mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]\,\tilde{V}_i^{-1}\right)$ *and* $\tilde{V}_i := \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}] + \frac{1}{\rho_i}\frac{dp}{dq_i}$.

Notice that if $dp/dq_i = 0$, then $\frac{\hat{V}_i}{2} = \tilde{V}_i = \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]$. This becomes Lemma 1.

This formula explains another important features of our results. Multiplying dp/dq_i is an investor's risk tolerance $1/\rho_i$. Since this is absolute risk aversion and we know that absolute risk aversion declines in wealth, one can interpret this as a proxy for investor wealth. In equilibrium, an investor with lower absolute risk aversion will have larger trade sizes, and their equilibrium trades will have more price impact.

The price impact of *all* investors' trades would seem to matter for the value of data. It does. But once again, it is captured by the variances. Of course, if the investor is large, it is possible that knowledge of this data choice will change the behavior of other investors – we abstract from this possibility, by positing a surprising, one-time deviation.

The two key assumptions behind both the competitive and market power results are that price can be approximated as a linear function of innovations as in Equation (4), and that individual i maximizes risk-adjusted return. In other words, this calculation is accurate as long as investors use linear factor models and maximize risk-adjusted return, even with potentially heterogeneous prices of risk.

Why it Doesn't Matter if Data is Public or Private. The expression for data value is the same for public and private information – regardless of who else knows the data, it is valuable only for its ability to change the conditional forecast errors. This might seem to contradict what we know about information value, e.g., Glode, Green, and Lowery (2012). The reconciliation comes from the fact that the publicity of the data does matter for the conditional variance. Private information, which is less likely to be impounded into price, is typically more valuable compared to information that the market already knows (and is therefore uncorrelated with $p_{t+1} + d_{t+1}$). Public information about $p_{t+1} + d_{t+1}$ is already impounded in rp_t.

In short, knowing the forecast errors fully captures the way in which knowledge matters: conditional variances, or in other words, the properties of forecast errors, are sufficient statistics. This is an incredibly helpful property because it relieves the econometrician of having to figure out who knows what.

Similarly, the risk preferences and investment styles of all market participants matter for the value of data. However, the expected profit $\mathbb{E}\left[\Pi_{it}\right]$ captures the way in which risk preferences and investment mandates matter.

Wealth Effects: Mapping Utility to a Dollar Value The dollar value of data is the amount of risk-free return an investor would require to be indifferent between having the data, or not having the data but getting the additional riskless wealth. Dividing the difference in utility by the coefficient of absolute risk aversion delivers a certainty equivalent amount:

$$\$\text{Value of Data}_i = \frac{1}{\rho_i}\left(\tilde{U}(\mathcal{I}_{it} \cup \text{data}) - \tilde{U}(\mathcal{I}_{it})\right) \tag{8}$$

This coefficient of absolute risk aversion is not constant (not CARA). It varies with wealth. Every utility function has some absolute risk aversion at every point of the function. It is this local, wealth-dependent risk aversion that we call ρ_i. One way to impute such a

value is to assume the investor has constant relative risk aversion (CRRA, denoted σ_{CRRA}).[7] Then, compute the level of absolute risk aversion that equates the two utilities, for each level of wealth: $-\exp(-\rho_i c_{it}) = c_{it}^{1-\sigma_{CRRA}}/(1-\sigma_{CRRA})$. For relative risk aversion of 2, we can express ρ_i as a function of investor wealth c_{it} as

$$\rho_i = \ln(c_{it})/c_{it}. \tag{9}$$

Data About Order Flow or Sentiment Many new data sources teach us about sentiment – something unrelated to the fundamental asset value, that affects current demand. Analyzing Twitter or StockTwits is one example. In our model, the variable that moves current price in a way that is orthogonal to value is x_{t+1}. So, we interpret sentiment as something that shows up in x. Farboodi and Veldkamp (2020) shows that such data can be used to remove the noise from price signals. Doing this is functionally equivalent to trading against dumb money, a common practice for sophisticated traders. Our sufficient statistics in Equations (6) and (8) can also value data series about sentiment, order flow, or aspects of demand that are orthogonal to future cash flows (see Appendix F).

2 Data and Estimation Procedure

Next, we describe our estimation procedure and the various data series used for returns and financial signals.

[7] An alternative approach to estimating ρ could be to use the market price of risk. Using the formulas for the equilibrium price coefficients, one could map the value of ρ to an equity premium and choose the value that matches a preferred estimate of the equity premium. We do not follow that approach for two main reasons. First, this would give us an estimate of the market's risk aversion and therefore, on how an average investor in the market values data. We are interested in how an individual investor, with particular characteristics should value data and in understanding how investor heterogeneity matters for data valuation. Second, it requires estimating most of the structural parameters of the model. As such, the estimates becomes much more sensitive. It negates the advantage of our sufficient statistics approach.

2.1 Data Sources

Our toolkit can used to value any finance-relevant data stream or bundle of data streams. In the rest of the paper, we show how it can be used to value two different data streams. The first, discussed in the following section, values earnings forecast data put out by stock analysts. We discuss how the value varies with investor heterogeneity along various dimensions and market conditions. The second, in Section 4, estimates the value of a hypothetical data source that allows investors to perfectly forecast GDP. Before turning to that analysis, we describe the two data sources of interest in more detail.

The Financial Data Stream We Value: I/B/E/S Forecasts The data series of interest in our first exercise is earnings forecasts provided by the Institutional Brokers' Estimate System (I/B/E/S). We use earnings forecasts for 12,501 unique firms from 1985–2019, with 2,597 firm observations per quarter on average.[8]

We use quarterly earnings forecasts from I/B/E/S. In our baseline model, investors have a horizon of a quarter and use the latest available one-quarter-ahead earnings forecast at each date. Later, we explore how different trading horizons affect the data value.

The Macro Data Stream We Value: *Ex-post* GDP Growth For realized GDP growth, we use the second release estimates of quarterly GDP growth from BEA, as reported by the Federal Reserve Bank of Philadelphia[9].

Data Sources for Asset Prices and Cashflows All data are for the U.S. equity market, over the period 1985–2019. Stock prices come from CRSP (Center for Research in Security Prices). All accounting variables are from Compustat. For our quarterly calculations, we use the market capitalization at the end of the calendar quarter and total dividends paid

[8]We use the Summary Statistics series from I/B/E/S, accessed through WRDS, `https://wrds-www.wharton.upenn.edu/pages/get-data/ibes-thomson-reuters/ibes-academic/summary-history/summary-statistics/`.

[9]`https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/routput`

throughout the calendar quarter. For annual calculations, we use market capitalization at the end of the calendar year and total dividends paid throughout the calendar year. In line with common practice, we exclude firms in the finance industry (SIC code 6).

We make a couple of adjustments to the raw data. First, we adjust for stock splits, buybacks and other events which affect shares outstanding for any stock, using the standard CRSP adjustment factors. The second pertains to exiting firms. Our preferred solution is to only consider periods during which a firm has non-missing information. Next, we winsorize the nominal values for market capitalization and total dividends at 0.01% level.

The equity valuation measure, i.e., the empirical counterpart for the price p_{jt} in the model, is market capitalization over total shares outstanding. Our cash-flow variable, d_{jt}, is total dividends paid over shares outstanding.

We use the holding period returns from CRSP to calculate the the excess returns using the yield on Treasury bonds (constant maturity rate, hereafter CMT) as the risk-free rate. For the quarterly estimates, we use CMT with three months maturity, while for annual estimates, we use CMT with one year maturity.

Forming Asset Portfolios The procedure described above can be used for any number and type of assets, including individual stocks. However, for expositional purposes, and to show more clearly the patterns in data value, we group assets into a small number of commonly-used portfolios, rather than work with a large number of individual stocks/assets. This leaves us with a more manageable number of data values to compute and compare.

Our first two portfolios are based on size. We group firms into Large and Small, based on whether their market capitalization is above or below the median value for all firms in our sample, in a given period. Next, we construct Growth and Value portfolios, using the book-to-market ratio (defined as the difference between total assets and long-term debt, divided by the firm's market capitalization). Firms above the median value of book-to-market in a period are assigned to the Value portfolio, while those below the median are part of the Growth portfolio. In addition to these four portfolios—Small, Large, Growth and Value—we

also include a market index (specifically, the S&P500) as a portfolio. We use value-weighted averages for excess returns for each portfolio as the return measure, where we weigh each firm's return by its market capitalization.

Measuring Price Impact Quantifying data value using (12) requires an estimate for price impact $\frac{dp}{dq_i}$. In practice, an investor who wants to value data using this toolkit should use a number appropriate to their context (i.e., how much the price of the asset typically moves when they trade). Estimates in the literature span a wide range—see Gabaix and Koijen (2021). In our baseline analysis, we will use one of those estimates, specifically the one from Frazzini, Israel, and Moskowitz (2018) who find that trading 2.5% of the daily volume of a stock has a price impact of 15 bp on the price.[10].

To map this estimate to an elasticity $\frac{dp}{p}/\frac{dq}{q} = 15$, we follow Gabaix and Koijen (2021) and assume an annual turnover of 100% and 250 trading days per year. The object $(dp/dq_i) \oslash p_t p_t'$ in our model can then be obtained by simply dividing this elasticity by the market capitalization pq. We use a reference market capitalization of USD 1 Billion for our exercises, which leads to $\lambda \equiv \frac{dp}{dq_i} \oslash \theta_i p_t p_t' \theta_i' = 1.5 \times 10^{-8}$. Data limitations force us to make the following simplifying assumptions: (i) price impact is the same for all portfolios we analyze (all θ_i's) and (ii) there is no cross-asset price impact, i.e. trading in one asset only move the prices of that asset. These assumptions can be relaxed, as per the appropriate market structure being studied for valuing relevant financial information. Under these two assumptions, the matrix $(dp/dq_i) \oslash \theta_i p_t p_t'$ in Equation (12) takes the form λI, where $\lambda = 1.5 \times 10^{-8}$ is the price impact and I is the identity matrix. While this value of λ might seem like a small number, we will see that it has substantial impact on data valuations.

[10]In Appendix G.2, we also report our data values from our baseline exercise using the price impact estimate of Gabaix and Koijen (2021).

2.2 Estimation Procedure

Excess Returns To build a tighter connection with the asset pricing literature, we reformulate our data value expression in terms of returns. Excess return on assets in the investment set is defined as:

$$R_{it} \coloneqq \theta_i \left[(p_{t+1} + d_{t+1}) \oslash p_t - r \right] = \Pi_{it} \oslash \theta_i p_t, \tag{10}$$

where $\oslash$ represents the Hadamard (element-by-element) division of two matrices. The binary θ_i matrix pre-multiplying returns selects out only the subset of returns that are for assets the investor can hold, given their investment style constraint. This ensures that investors do not get expected utility from assets they cannot hold, and drops out ($\theta_i = I$) for investors who trade all assets.

The investor unconditional expected utility in Lemma 1 and Lemma 2 are expressed in terms of Π_{it}. In Appendix E, we derive expressions for *ex-ante* expected utility expressions in Lemma 1 and Lemma 2 in terms of moments of returns.[11] In the case of perfect competition ($n \to \infty$), expected utility is

$$\tilde{U}(\mathcal{I}_{it}) \approx \frac{1}{2} \left\{ \mathbb{E}\left[R_{it}\right]' \mathbb{E}\left[\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1} \right] \mathbb{E}\left[R_{it}\right] \right\} + \frac{1}{2} \operatorname{Tr}\left[\mathbb{V}\left[R_{it}\right] \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1} - I \right] + r \overline{w}_{it} \rho_i. \tag{11}$$

If investors have price impact, expected utility is

$$\tilde{U}(\mathcal{I}_{it}) \approx \mathbb{E}\left[R_{it}\right]' \hat{\tilde{V}}_{it}^{-1} \mathbb{E}\left[R_{it}\right] + \operatorname{Tr}\left[\left(\mathbb{V}\left[R_{it}\right] - \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] \right) \hat{\tilde{V}}_{it}^{-1} \right] + r \rho_i \bar{w}_{it}, \tag{12}$$

where $\hat{\tilde{V}}_{it} \coloneqq \tilde{\tilde{V}}_{it} \left(I - \frac{1}{2} \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] \tilde{\tilde{V}}_{it}^{-1} \right)^{-1}$ and $\tilde{\tilde{V}}_{it} \coloneqq \left(\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i} \frac{dp}{dq_i} \oslash \theta_i p_t p_t' \theta_i' \right)$.

[11] This requires an assumption about the *ex-ante* variability of p_t. The key patterns in data valuation described in the following sections hold even when we work directly with profits using the expressions in Lemma 1 and Lemma 2.

Empirical Specification The first step is to construct a time series of the return vector R_t by computing returns for each asset type, R_{jt}. The estimates for unconditional expected return $\mathbb{E}[R_t]$ and variance $\mathbb{V}[R_t]$ are obtained from the corresponding time series moments, i.e., $\widehat{\mathbb{E}[R_t]} = \frac{1}{T}\sum_{t=1}^{T} R_t$ and $\widehat{\mathbb{V}[R_t]}$ is the Newey-West estimator for the spectrum at frequency zero of the series $R_t - \widehat{\mathbb{E}[R_t]}$, with the Andrews rule for HAR inference bandwidth $0.75T^{\frac{1}{3}}$ (Andrews, 1991).

Our strategy requires a historical time series of the data-set they are interested in valuing. The next step is to project R_t on the available time series of the data along with any other data that the investor already has access to. In our baseline empirical implementation, we will use a standard variable, namely the consumption-wealth ratio (cay) from Lettau and Ludvigson (2001), as a proxy for such existing data. The procedure is a ordinary least squares regression of returns R_t on all the variables, already owned and new, in the data set. The estimated variance of the residuals is then our estimate for $\mathbb{V}[R_t \mid \mathcal{I}_{it}]$.

Using these objects, we can compute $\mathbb{E}[U(c_{it+1})]$. We then repeat this procedure excluding the data series of interest, i.e., with only the already-owned data. The difference between these two expected utilities is the utility gain from having access to that data source.

Formally, given data, denoted X_t, and existing data, denoted Z_t, we can estimate the precisions $\mathbb{V}[R_t \mid X_t, Z_t]^{-1}$ and $\mathbb{V}[R_t \mid Z_t]^{-1}$ by estimating the following two regressions:

$$R_t = \beta_1 X_t + \beta_2 Z_t + \varepsilon_t^{XZ} \tag{13}$$

$$R_t = \gamma_2 Z_t + \varepsilon_t^{Z} \tag{14}$$

For a single-asset case, the two regression specifications in Equations (13) and (14) are estimated through OLS. The data X_t indicates the data signal we are interested in valuing, and Z_t is the set controls which are already present in the investor's information set. In case R_t has more than one asset's return (that is, the investor's investable universe consists of more than one asset), the regression equations form a seemingly unrelated regression

(SUR) system. For the purposes of our exposition, we assume that the data series being valued X_t is the same for all assets, and the set of controls are also the same across all assets in this SUR system. Thus, since the independent variables for each asset in this SUR system are the same, equation-by-equation OLS can again be used to efficiently estimate the system (Baltagi, 2021). From these two vector regressions, an estimate for $\mathbb{V}\left[R_t \mid \mathcal{I}_{it}\right]$ would be $\widehat{\text{Cov}}(\varepsilon_t^{XZ})$. Similarly, the estimate for $\mathbb{V}\left[R_t\right]$ would be $\widehat{\text{Cov}}(\varepsilon_t^{Z})$. We calculate both these conditional variances using the Newey-West estimators of the spectrum at frequency zero (similar to $\widehat{\mathbb{V}\left[R_t\right]}$) for the residuals ε_t^{XZ} and ε_t^{Z}, respectively. Substituting in the mean return and the estimated variance-covariance matrices in Equation (6) yields the estimated value of data, in utils. To get the standard errors for the estimated data value in each case, we use a wild bootstrap with 500 samples using Equation (14) as the DGP, and the two-point distribution of Mammen (1993).

One might question how a Bayesian theory corresponds to a procedure that uses OLS. When variables are normal and relationships are linear, Bayesian estimates are the efficient, unbiased estimates. Since OLS estimates are the unique efficient, unbiased linear estimates, they must coincide with the Bayesian ones, in the specific case of normal variables in a linear relationship. Thus, in this case, OLS estimators are Bayesian weights on information. In cases where variables are not normal or the expected relationship between the data and R_t is not linear, there are a few possible solutions: 1) Transform the data to make it normal or linear; 2) use OLS or non-linear least squares as an approximation to the Bayesian forecast, or 3) perform a Bayesian estimation.

Data Timing As discussed above, our return measure for year t for an asset j is the cum-dividend excess return on that asset over the year t – using prices at the end of year t and at the end of year $t-1$, along with dividends paid out over year t. We are interested in understanding the value of data available to an investor *before* year t, in predicting the value of this return measure for year t.

The value of any control variable in Z_t used for the purpose of this calculation is obtained

for year $t-1$, since these values will be in the investor's information set while predicting the profits for year t. Similarly, the data signal in X_t that we are valuing needs to be in the information set of the investor *before* year t. To predict profits over year t, we use the data signals which are produced *before* year t starts, which give information about growth in earnings of firms between year $t-1$ and year t.

3 Valuing Financial Data

In this section, we first estimate the utility gain that investors would assign to I/B/E/S forecasts, given what they already know, and then convert this into a dollar amount. The latter is the monetary value of I/B/E/S data, or equivalently investors' willingness to pay for this data. In most cases, these private valuations look nothing like a price that any investor actually pays for an I/B/E/S subscription. Some valuations are orders of magnitude higher, others much lower. Recall that these are not predicted transactions prices. They are private valuations that trace out a demand curve. The qualitative patterns are mostly intuitive, which suggests that our measurement strategy/toolkit is a sensible one.

When we value a stream of data, we need to take a stand on what else an investor already knows, i.e. the publicly available information. Obviously, as econometricians, we do not observe information sets directly, so in our implementation, cannot control for this perfectly. Of course, this is not a problem for a practitioner or investor who wishes to use our toolkit to value a stream of data (e.g. one that she is considering buying), since she would know exactly what other data she already has access to. For the purposes of illustrating the use of the tool, we endow our hypothetical investor with some commonly-used and publicly-available data series. Specifically, we assume that they already observe the consumption-wealth ratio (cay) from Lettau and Ludvigson (2001). In addition to this control variable, we also assume that the investor observes the realized price p_t for the asset they are trading in, as an additional control variable. Since investors in our quantitative

exercises trade in portfolios, we use the value-weighted mean price for the corresponding portfolio. For the case of an investor trading in multiple portfolios, we provide the price corresponding to the S&P500 portfolio as the additional price control.

In additional results, we also consider an investor who also has access to one or more of the following pieces of data: the S&P500 dividend yield (D/P ratio)[12], the yield on a 1-year Treasury bill (constant maturity rate)[13], and a sentiment index from Baker and Wurgler (2006).

We use quarterly earnings forecasts from I/B/E/S as the first data stream that we value. In our baseline model, investors have a horizon of a quarter and use the latest available one-quarter-ahead earnings forecast at each date. We use annualized return moments for all data value calculations, and report the dollar values in annualized thousands of USD.

For each firm, we use the median consensus analyst forecast for earnings per share (hereafter EPS). We discard all forecast values which have been calculated during or after the period for which the forecast is being made. For example, any forecast we use for earnings in 2015 Q1 has to be issued before 2015 Q1 starts. We then drop all but the latest consensus forecasts for each firm–period observation, which gives us a single consensus forecast for EPS over the next period. Using this forecast, we calculate a forecasted growth rate: the forecasted EPS for the coming period, divided by the realized value of EPS from the last period.

Our goal is to explore data valuation patterns, to get a sense of how large this amount is and to gain some intuition about what makes it vary. In order to keep the analysis manageable, we collapse the large number of assets into a few portfolios. Specifically, we analyze five portfolios: Small, Large, Growth, and Value firms, as well as the S&P500 index.

We find that most of the value of the I/B/E/S data comes from earnings forecasts of growth firms and those in the S&P500 index. Therefore, the data value numbers we report

[12]Obtained from NASDAQ Quandl `https://data.nasdaq.com/data/MULTPL/SP500_DIV_YIELD_MONTH-sp-500-dividend-yield-by-month`.

[13]Obtained from FRED series `DGS1`.

in this section for two quarterly signals, one about earnings of all firms in the Growth portfolio and one about the earnings of all firms in the S&P500 index. Specifically, these are the portfolio value-weighted average values of median forecasted growth rates for earnings per share—for the Growth and S&P500 portfolios. Note that we are valuing a forecast of a payoff of a particular portfolio of assets. [14]

3.1 Wealth and Risk Tolerances

One obvious dimension along which investors differ is the size of their portfolios. We consider investors with two wealth levels—$\$500,000$ and \$250 million, each with the same relative risk aversion of $\sigma = 2$. In terms of the wealth level, the former group is similar in magnitude to the wealth level of the mean US household (Badarinza, Campbell, and Ramadorai, 2016), while the latter investor group has comparable wealth level to the size of the mean US hedge fund (Yin, 2016). The resulting difference in absolute risk aversion give rise to different willingness to pay for the same data.

To value data for a particular investor, we need to know what else they already know and what they can invest in. Our investor are assumed to know the consumption-wealth ratio (cay) and the value-weighted mean price for the S&P500 portfolio at the end of the previous period. They can invest in any combination of the following five portfolios: S&P500, Small, Large, Growth, and Value. However, we make no assumption about what any other investors know or trade.

Table 1 reports the dollar value of the I/B/E/S forecasts for two investors with different wealth levels, with and without price impact, who can invest in five portfolios: {Small, Large, Growth, Value, S&P500}. The results illustrate that wealthier investors attach a

[14] We could have performed this calculation under many alternative assumptions. For example, one could value growth firms' data from the perspective of an investor who invests only in growth firms. In that case, one would regress the growth firm asset payoffs on the relevant data and use means variances and forecast errors of growth asset payoffs. We did not take that approach because if we vary the investment set and the data together, we would not know whether data was more/less valuable because of the data or the investment restriction. But, it is certainly another dimension of investor heterogeneity that might be interesting to explore.

Table 1: **Risk Tolerance, Liquidity and Data Value.** The table reports the valuation of the quarterly value-weighted means of I/B/E/S median forecast earnings growth for Growth and S&P500 portfolios. Quarterly data between 1985–2019. The dependent variables in (13) and (14) are the vector of returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. The specification includes a constant and control variables (cay and value-weighted mean price of the S&P500 portfolio). The case with price impact assumes Kyle Lambda $\lambda = 1.5 \times 10^{-8}$. Absolute risk aversion is from (9). All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Perfect Competition	With Price Impact
Panel A: Investor with $500,000 Wealth.		
Utility Gain	0.082	0.053
	(0.02)	(0.02)
Expected Profit	0.008	0.007
	(0.01)	(0.01)
Variance Reduction	0.075	0.046
	(0.013)	(0.013)
Dollar Value (in $000)	3.13	2.03
	(0.75)	(0.75)
Panel B: Investor with $250m Wealth.		
Utility Gain	0.082	0.01
	(0.02)	(0.005)
Expected Profit	0.008	0.001
	(0.01)	(0.001)
Variance Reduction	0.075	0.009
	(0.013)	(0.005)
Dollar Value (in $000)	1062.69	128.96
	(254.07)	(66.08)
Time Periods	140	140

higher dollar value to the same data. In our setting, this occurs through the dependence of the curvature parameter ρ on wealth. Under our calibration, an investor with $250 million in wealth operating in a competitive setting would be willing pay almost 340 times more for this data compared to an investor with half a million dollars of wealth. Also note that while investors with different wealth levels experience the same utility gain with the data stream under the perfectly competitive case, they still assign different dollar values owing to the wealth effects affecting their local risk aversion. However, when we add a plausible realistic

degree of price impact, wealth effects show up directly in utility gains as well, not just in dollar values.

Next, as one would expect, price impact attenuates the value of data. This is intuitive: investors use the data that they acquire to trade more profitably. When they face price impact, they cannot incorporate the data in their trading strategy as effectively as they move the prices against themselves, which in turn implies a decline in their data valuation. The table shows that this effect is quite significant and increases with wealth. To see why, recall that in Lemma 2, price impact (dp/dq) gets scaled by $1/\rho_i$. Since wealthier investors are assumed to have a lower degree of absolute risk aversion (a lower ρ_i), price impact has a disproportionate effect on their payoffs and data valuations. For an investor with \$250 million in wealth, taking price impact into account cuts the value of the I/B/E/S data by 90%.

To better understand the sources of data value, Table 1 also reports the expected return and the variance reduction on the investor's portfolio. The expected profit is the ex-ante expected return on the optimal, diversified portfolio of the five assets the investor can hold. The variance reduction is the difference between the raw variance of this return and the conditional variance, which is the average squared residual of the predicted return, after conditioning on the data. This is a measure of how much one learns from data. Notice that price impact lowers both components of data value and has a more pronounced effect when wealth is higher (or equivalently, absolute risk aversion is lower).

3.2 Investment Styles

Investors also differ in their investment style. To understand the implications of this type of heterogeneity for data valuation, we value exactly the same data as before, the median earnings growth forecasts from I/B/E/S, from the perspective of investors who only trade in individual portfolios. We will refer to these investors by the portfolios they trade. For example, the Value investor is one who only buys and sells the portfolio of Value stocks that

Table 2: **Investment Styles and Data Value.** Value of I/B/E/S data (quarterly value-weighted means of I/B/E/S median forecasted earnings growth for Growth and S&P500 portfolios). Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant and control variables (cay and value-weighted mean price of the corresponding portfolio). In the last column, we control for a constant, cay and the value-weighted mean price of the S&P500 portfolio. The price impact panel assumes Kyle Lambda $\lambda = 1.5 \times 10^{-8}$. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Investment Style					
	Small	Large	Growth	Value	S&P500	Multi-Asset
Panel A: Perfect Competition.						
Dollar Value (in $000) for	0.0	0.33	0.34	0.15	1.0	3.13
Investor with $500,000 Wealth	(·)	(0.24)	(0.28)	(0.25)	(0.46)	(0.75)
Dollar Value (in $000) for	0.0	110.9	116.42	52.27	340.22	1062.69
Investor with $250m Wealth	(·)	(81.11)	(96.31)	(86.14)	(157.62)	(254.07)
Panel B: With Price Impact.						
Dollar Value (in $000) for	0.0	0.33	0.34	0.15	1.0	2.03
Investor with $500,000 Wealth	(·)	(0.24)	(0.28)	(0.25)	(0.46)	(0.75)
Dollar Value (in $000) for	0.0	53.86	58.48	11.33	56.35	128.96
Investor with $250m Wealth	(·)	(38.21)	(46.18)	(18.83)	(30.15)	(66.08)
Time Periods	140	140	140	140	140	140

we constructed. They each use the earnings forecast data to determine how much to trade in their respective portfolios. We compare these data values to the value of the investor who trades in all five portfolios (Small, Large, Growth, Value and S&P500), which corresponds to the case analyzed in Table 1.

Table 2 shows that among the investors who invest in a single portfolio, I/B/E/S forecast data is most valuable for investors in Growth, Large or the S&P500 portfolios. While the investor wealth and price impact raise and lower the dollar value of the data, respectively, this pattern of Growth, Large, and S&P500 investors valuing earnings forecast data by more emerges consistently.[15] This is because the I/B/E/S data lacks relevance for the Small

[15]We report the unconditional and conditional moments for the case with perfect competition in Appendix G.3.

portfolio. More precisely, it does little to reduce return forecast errors and in that sense, provide little guidance to an investor about when to buy and sell the Small portfolio. So, despite the high unconditional expected returns of the Small portfolio, the value of this particular data stream for such investors is quite low. The relevance of the I/B/E/S forecasts is low for the Value portfolio as well. The Large and Growth portfolios on the other hand have medium expected returns, but their returns are predicted to a larger degree by the analyst forecast data. Therefore, this data is most valuable to those who invest in equity portfolios consisting of growth and large firms.

As we saw in the previous set of results, price impact reduces the value of data, but also reduces the dispersion in valuations. The investors who value data most are the same investors who would like to trade aggressively on the data, but are prevented from doing so when price impact is large.

It is worth re-iterating that our approach estimates the value of data in an equilibrium setting, where prices are noisy signals of asset performance. Our regression specifications explicitly control for the information contained in market prices (by including the price p_t in the investor's information set). In Appendix G.1, we conduct a counterfactual exercise—where investors are assumed to not learn from prices—to tease out the effect of equilibrium learning on our data value estimates. We find that incorporating price information changes data valuations meaningfully, demonstrating the importance of equilibrium forces. It is possible for price information to increase the value of data in a setting when price noise and signal noise are negatively correlated (see Appendix H).

Data and Diversification In this set of results, data is always most valuable to the multi-asset investor. This investor can use data not only to decide whether to buy more or less at a given moment in time, but also to decide what to buy. The multi-asset investor can use data for asset allocation. However, it is possible for the multi-asset investor to value data

less.[16] The reason this arises in some cases is that the multi-asset investor can diversify. This investor has two tools to reduce risk: information and diversification. Each is an imperfect substitute for the other. But the ability to diversify depresses the value of data. In this case, the increase in value from asset allocation greatly outweighs the decrease in value from the ability to diversify.

3.3 Market Liquidity

A consistent theme throughout our results is the importance of price impact. For expositional purposes, we have treated price impact as a single, time-invariant number. In reality, it fluctuates with market liquidity. Our estimates suggest that such fluctuations will have a dramatic impact on the value of data, especially for large investors.

Now consider a financial firm whose business model revolves around the use or sale of data. That firm's market value is based largely on the value of their data. Changes in market liquidity will thus affect the real value of this firm's data assets through this channel.

As firms' data stocks grow larger, the effect of liquidity shocks on data values should grow. The reason is that price impact enters additively with conditional variance. This additive form comes from first order condition for the optimal portfolio choice of investor i: $q_{it} = (\rho_i \mathbb{V}\left[p_{t+1} + d_{t+1} | \mathcal{I}_{it}\right] + dp/dq_i)^{-1} (\mathbb{E}\left[p_{t+1} + d_{t+1} | \mathcal{I}_{it}\right] - r p_t)$. If the conditional variance $\mathbb{V}[p_{t+1} + d_{t+1} | \mathcal{I}_{it}]$ is large (high uncertainty), then small changes in price impact dp/dq_i have little effect. Those changes are swamped by the variance term and the inverse of this large number is small. However, if conditional variance is small, meaning that return forecasts are relatively precise, then that first term, the inverse of a potentially small number, may be large. In this case, the effects of price impact can be substantial. Over time, if firms have more data and thus smaller forecast errors, their data valuations become more and more

[16] Consider a simple example to illustrate this point: an economy with two assets, and a simple factor structure in their payoffs. Asset 1 pays $f_1 + f_2$, while Asset 2 pays $f_1 - f_2$. The factors f_1 and f_2 are orthogonal. Suppose the equilibrium portfolio of a multi-asset investor has equal holdings of each asset. Such an investor does not value information about the factor f_2 at all, even though a single asset investor in either asset would value such information non-trivially.

susceptible to changes in the price impact of a trade.

Our estimation results illustrate that price impact significantly reduces the value of data for all investors irrespective of their wealth, risk tolerance, and investment style. Furthermore, incorporating price impact uncovers a novel insight: inelastic asset demand can be accompanied with more elastic data demand. On the one hand, price impact (illiquidity) causes investors to reduce the sensitivity of their trading decisions to prices and leads to a low price elasticity of asset demand. On the other hand, price impact makes data valuations less heterogeneous by lowering data value most significantly for investors with the highest data valuations to begin with, leading to a high price elasticity of data demand.

The high and growing sensitivity of data value to market liquidity suggests a new channel through which market liquidity matters. Since the value of a financial firm depends on its ability to trade profitably, the value of data is an input into the valuation of a financial firm. As financial firms become more data-centric, the firm's value becomes more sensitive to the value of its data. At the same time, growing data abundance makes the value of data more sensitive to market liquidity. These two margins of increasing sensitivity amplify each other. This suggests that changes in market liquidity may affect the real value and the equity value of financial firms through a new channel, through the value of their data. In a world in which data is becoming increasingly abundant, this new liquidity-data effect could grow much stronger. These findings suggest that, because of the rising abundance and importance of data for financial firms, market liquidity may become more important than ever before.

3.4 Previously Purchased Data

A third dimension along which investors differ enormously is in the data they already own. While large, institutional investors have access to enormous libraries of data, households may know only a few summary statistics about each asset. Here, we present a few exercises to explore the effect of this dimension of heterogeneity. In our baseline exercise, we valued

Table 3: **Previously Purchased Data.** Values in each row represent the additional value of I/B/E/S data (specifically, the quarterly value-weighted means of I/B/E/S median forecasted earnings growth for Growth and S&P500 portfolios) *on top of* the value-weighted mean price of the S&P500 portfolio, and control variable(s) listed in the first column, with price impact assuming Kyle Lambda $\lambda = 1.5 \times 10^{-8}$. Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT) for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

Other Data	Dollar Value (in $000) for Investor with	
	Wealth: $500,000	Wealth: $250m
cay (Baseline)	2.03	128.96
	(0.75)	(66.08)
No Other Data	2.72	119.73
	(0.72)	(66.54)
Real CMT-1yr	2.09	142.26
	(0.75)	(66.47)
S&P500 D/P ratio	1.65	117.78
	(0.7)	(58.86)
BW Sentiment Index	3.74	234.61
	(0.86)	(75.97)
All	1.51	113.48
	(0.8)	(63.71)
Time Periods	140	140

the I/B/E/S data assuming that investors also access to a common variable used to predict returns, namely the consumption-wealth ratio (cay) from Lettau and Ludvigson (2001). We now ask: How valuable would the same I/B/E/S forecasts be if, instead of the cay series, the investor had some other variable in his or her existing data set? Of course, this does not nearly capture the extent of the difference between the knowledge of investors. But, it will help shed light on the sensitivity of data values to other sources of information.

In Table 3, the first row reports our baseline estimates—from Table 1—for the value of the I/B/E/S forecasts to investors with different wealth levels who trade all five portfolios and have price impact. The second row 'No Other Data' shows the value if the investor does not have access to the cay series. The remaining rows report the value of the same

data stream for investors who have access to other data series (instead of cay): specifically, S&P500 D/P ratio, Real CMT-1yr and the investor sentiment index from Baker and Wurgler (2006). Finally, the last row assumes that the investor has access to all of these data series (as well as an additional macro variable, realized year-on-year inflation).

The analysis yields two insights. First, differences in other data available to the investor can induce substantial variation in the valuation of a given data series. An investor who has (only) the Baker-Wurgler sentiment index in his information set values the I/B/E/S data more than twice as much as an investor with access to all the other data series. Second, the I/B/E/S forecasts remain valuable even for a relatively sophisticated investor. For example, our valuation estimate for an investor with $250 million in wealth who uses all the data series mentioned in the table is $113,000, which is only 12% lower than the baseline case (where the investor only used the cay series). This suggests that the information contained in the I/B/E/S data cannot be easily substituted with other aggregate data.

3.5 Trading Horizon

Finally, investors differ in their trading horizons. Our data valuation tool can be applied to various trading horizons. However, for the data we are exploring, this dimension of investor heterogeneity seems to matter less than the others. Our calculations so far have assumed that investors trade over a quarterly horizon. Next, we measure the value of the data series with an annual horizon—the median I/B/E/S forecast for earnings over the next one year—for an investor who trades the same portfolio but with an annual horizon. This does not change the data value formula; it does change how we implement it. The procedure is to compute residuals from (13) and (14) where R_t is annual return, the prior information Z_t is a constant and cay, and where X_t is the median forecast of the earnings growth for Growth and S&P500 portfolios over the next year. The resulting regression residuals (ε_t^{XZ} and ε_t^{Z}) are then used to construct the variance matrices and substitute these variances, along with expected annual returns, into the expected utility formula (6). We convert expected utility

Table 4: **Trading Horizon.** Data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant, cay, and price of S&P500 portfolio. Data variables being valued are the value-weighted means of I/B/E/S median forecasted annual earnings growth, reported quarterly, for Growth and S&P500 portfolios. Data variables correspond to earnings forecasts with corresponding time horizon. Values are calculated with price impact assuming Kyle Lambda $\lambda = 1.5 \times 10^{-8}$. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Quarterly	Annual
Dollar Value (in $000) for	2.03	20.45
Investor with $500,000 Wealth	(0.75)	(4.86)
Dollar Value (in $000) for	128.96	198.73
Investor with $250m Wealth	(66.08)	(184.07)
Time Periods	140	35

to data value as before, using (8). Since we use annualized return moments to obtain the data value throughout our quantification exercises, the data value estimates are directly comparable for investors across both trading horizons.

Table 4 reports the value of the I/B/E/S forecasts for both annual and quarterly investors. The first column is the same values reported in Table 1. The second column shows that investors who trade less frequently, on an annual basis, would be more willing to pay for similar data. The reason for the lower valuation for more frequent observations is that quarterly returns are considerably more noisy. Earnings data is not very useful for quarterly portfolio adjustment. Trading on this data only creates more noise.

The effect of trading horizon surely depends on the data source. For example, high-frequency data is useful for high-frequency traders, but will likely be worthless after a year. The more important take-away is that trading horizon can matter for how an investor values their data. By adjusting the input data and the interpretation of the results, our data valuation tool can be used to value data used by investors who trade at various frequencies.

3.6 Supporting Evidence

Finding direct evidence in support of our estimates is tricky. However, indirect support for our approach comes from evidence on investor data reflected in asset prices. If a certain kind of data is very valuable, then many investors should acquire it, and this information should be reflected in asset prices. If earnings data about a particular asset class (e.g. large firms) is more valuable (say, relative to that of small firms), then investors in that asset class should acquire more data and prices of large firm stocks should ultimately reflect more of this data than those of smaller firms.

Bai, Philippon, and Savov (2016) show that prices of S&P 500 firms incorporate more information over time, while other prices do not. This is consistent with column 5 of Table 2, showing that the value of data for an investor that holds the S&P 500 is about 3 times more valuable than for an investor who holds any other single portfolio. S&P investors should be acquiring more data. Farboodi, Matray, Veldkamp, and Venkateswaran (2019) break out the information component of the BPS measure, from the growth and volatility pieces, and show that large, growth firm data is becoming more abundant, while other types of data are not. This is consistent with the estimates from Table 2, showing that among the FF portfolios, large and growth investors value data more than small and value investors. Davila and Parlatore (2021) show that asset turnover is a significant predictor of asset price informativeness. To the extent that asset turnover is an indicator of market liquidity or lower price impact, then this finding supports a third key prediction of the model. This higher value of data about liquid assets shows up as more data acquired about such assets. In this sense, all three studies on price informativeness are consistent with the variation in data values that we estimate.

Table 5: **Macroeconomic Information.** Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant and controls for cay, value-weighted mean price of the corresponding portfolio, the realized real GDP growth in the previous quarter and the median forecasted quarterly growth rate in real GDP from the Survey of Professional Forecasters. In the last column the price control is for S&P500 portfolio. Data variables (X_t in (13)) are the second release estimates of real quarterly GDP numbers as reported by by the BEA, expressed as growth rates over the previous quarter. Numbers reported in each column represent the additional value of *ex-post* real GDP growth data (8) on top of the control variables for an investor trading at the quarterly horizon. The case with price impact assumes Kyle's Lambda $\lambda = 1.5 \times 10^{-8}$. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Investment Style					
	Small	Large	Growth	Value	S&P500	Multi-Asset
Panel A: Perfect Competition.						
Dollar Value (in $000) for	2.45	3.29	3.29	1.74	3.84	4.05
Investor with $500,000 Wealth	(0.4)	(0.42)	(0.44)	(0.4)	(0.41)	(0.99)
Dollar Value (in $000) for	830.01	1116.13	1117.42	589.57	1303.35	1374.21
Investor with $250m Wealth	(136.73)	(142.78)	(149.14)	(134.55)	(139.17)	(334.93)
Panel B: With Price Impact.						
Dollar Value (in $000) for	2.45	3.29	3.29	1.74	3.84	4.04
Investor with $500,000 Wealth	(0.4)	(0.42)	(0.44)	(0.4)	(0.41)	(0.99)
Dollar Value (in $000) for	276.59	504.29	522.44	126.19	201.83	836.43
Investor with $250m Wealth	(50.51)	(65.08)	(63.84)	(31.57)	(26.15)	(98.24)
Time Periods	140	140	140	140	140	140

4 Valuing Macroeconomic Information

How do different investors value information about macroeconomic variables (e.g. GDP)? We now use our framework to provide an answer to this question. In Table 5, we compute the value of a hypothetical data source which allows investors to perfectly forecast GDP. Formally, we use the realized (i.e. *ex-post*) real GDP growth as our data series of interest[17] and calculate its value to investors with different trading styles, defined in Section 3.2. We control for cay and the corresponding portfolio price, as before, as well as two addi-

[17]We use the second release revised estimates of realized real GDP growth for this calculation.

tional controls—the realized real GDP growth rate in the previous quarter and the median forecasted growth rate in real GDP for the current quarter by the Survey of Professional Forecasters (SPF)[18]. Adding these two additional control variables allows us to find the value of the new information in *ex-post* GDP growth.

The last column of Table 5 shows that a fund with assets of $250 million, trading all five portfolios under perfect competition, would be willing to pay $1.4 million for the ability to perfectly forecast quarterly GDP growth in advance. The other columns show values for more restricted trading styles. They are all sizable, albeit with some variation. There are some interesting cross-sectional differences relative to the value of earnings forecasts analyzed in Table 2. For example, better information about GDP is quite valuable for investors trading only the Small portfolio. This is because, unlike the earnings forecasts, GDP growth turns out to be a valuable predictor of returns on the Small portfolio, i.e., this data has high relevance for such an investor. In fact, macroeconomic information of this form shows relatively high data relevance for all five assets—unlike the earnings forecasts data, which showed high relevance mostly for Large and Growth portfolios.

The bottom panel shows the value of data with price impact. As with the earnings forecast data, price impact significantly attenuates the value of macroeconomic information as well and the effects are more pronounced for wealthier, less risk-averse investors. The value of a perfect GDP forecast for the aforementioned $250 million fund trading all five portfolios is cut more than half once price impact is taken into account. The drop in valuations is even more significant for some of the individual portfolios, again underscoring the importance of market liquidity for the value of data.

5 Conclusion

Data is one of the most valuable assets in the modern economy. Yet the tools we have to quantify that value are scant. We offer a tool that an investor or financial firm can use to

[18] `https://www.philadelphiafed.org/surveys-and-data/rgdp`

value its existing data, or a potential stream of data that it is considering to acquire. Along with information about the distribution of investor characteristics, researchers can use this tool to trace out the demand curve for data.

We uncover important investor wealth and trading style effects, the importance of an investor's existing data, and the role of trading horizon. Jointly, these effects point toward enormous heterogeneity, spanning multiple orders of magnitude, in the value different investors assign to the same data. The dispersion in valuations suggests that marginal changes in the price of data will have little effect on demand. With such dispersed valuations, few data customers would be on the margin. This low price elasticity of demand is significant because it points to one reason why data markets might not evolve to be very competitive.

We further uncover a new channel through which market liquidity matters for the real value of data, which is an important new class of assets. As firms accumulate more data and data technologies improve, more and more of the value of a financial firm will depends on the value of the data it possess. The sensitivity of the value of data to price impact of a trade could introduce a new source of financial fragility, brought on by data accumulation, and exacerbated by data technologies that improve financial forecasting.

The advantage of our measurement tool is its simplicity. While our measure of the value of data is derived from a structural model, computing it does not require estimating structural parameters. Instead, the relevant sufficient statistics are simple means and variances of linear regression residuals. No matter whether the data is public, private, or known only to a fraction of investors, these methods are valid. Even if the data is about sentiments or order flows, as long as it is measured along with the market prices in the observable data set, our data value measure offers a meaningful assessment of its value to an investor.

In order to make the paper most transparent, we used mean-variance preferences and a model without true dynamics. Appendix C offers an argument for using mean-variance preferences as a second order approximation for more general utility functions. One area of future work is to extend this analysis and provide bounds for our approximation errors.

Code Availability: The replication code is available in the Harvard Dataverse at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BDYCZ3`

References

Admati, A., and P. Pfleiderer (1986): "A monopolistic market for information," *Journal of Economic Theory*, 39(2), 400–438. 6

——— (1987): "Viable allocations of information in financial markets," *Journal of Economic Theory*, 43(1), 76–115. 6

Albagli, E., C. Hellwig, and A. Tsyvinski (2014): "Risk-Taking, Rent-Seeking, and Investment When Financial Markets Are Noisy," Yale Working Paper. 7

Amador, M., and P.-O. Weill (2010): "Learning from prices: Public communication and welfare," *Journal of Political Economy*, forthcoming. 7

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica: Journal of the Econometric Society*, pp. 817–858. 20

Badarinza, C., J. Y. Campbell, and T. Ramadorai (2016): "International Comparative Household Finance," *Annual Review of Economics*, 8(1), 111–144. 24

Bai, J., T. Philippon, and A. Savov (2016): "Have Financial Markets Become More Informative?," *Journal of Financial Economics*, 122 (35), 625–654. 7, 34

Baker, M., and J. Wurgler (2006): "Investor sentiment and the cross-section of stock returns," *Journal of Finance*, 61 (4), 1645–1680. 23, 32

Baltagi, B. (2021): *Econometric Analysis of Panel Data (Springer Texts in Business and Economics)*. Springer, 6th ed. 2021 edn. 21

Barlevy, G., and P. Veronesi (2000): "Information Acquisition in Financial Markets," *Review of Economic Studies*, 67(1), 79–90. 7

Bergemann, D., A. Bonatti, and A. Smolin (2016): "The Design and Price of Information," CEPR Discussion Papers 11412, C.E.P.R. Discussion Papers. 7

Campbell, J. Y., and S. B. Thompson (2008): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *The Review of Financial Studies*, 21(4), 1509–1531. 7

Davila, E., and C. Parlatore (2021): "Identifying Price Informativeness," NYU Working Paper. 7, 34

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(4), 877–909. 6

Farboodi, M., A. Matray, L. Veldkamp, and V. Venkateswaran (2019): "Where Has All the Data Gone?," Working Paper. 8, 34

Farboodi, M., and L. Veldkamp (2017): "Long Run Growth of Financial Technology," Working Paper, Princeton University. 7, 55

——— (2020): "Long-run Growth of Financial Data Technology," Discussion Paper 8. 15

Frazzini, A., R. Israel, and T. J. Moskowitz (2018): "Trading Costs," *Available at SSRN 3229719*. 18

Gabaix, X., and R. S. Koijen (2021): "In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis," Discussion paper, National Bureau of Economic Research. 18, 56

Glode, V., R. Green, and R. Lowery (2012): "Financial Expertise as an Arms Race," *Journal of Finance*, 67(5), 1723–1759. 14

Goldstein, I., E. Ozdenoren, and K. Yuan (2013): "Trading frenzies and their impact on real investment," *Journal of Financial Economics*, 109(2), 566–82. 7

Grossman, S., and J. Stiglitz (1980): "On the impossibility of informationally efficient markets," *American Economic Review*, 70(3), 393–408. 7

He, Z. (2009): "The Sale of Multiple Assets with Private Information," *Review of Financial Studies*, 22, 4787–4820. 7

Kacperczyk, M., J. Nosal, and S. Sundaresan (2021): "Market Power and Informational Efficiency," Working Paper, Imperial College London. 2, 6, 7

Kadan, O., and A. Manela (2019): "Estimating the Value of Information," *Review of Financial Studies*, 32 (3), 951–990. 6

Koijen, R. S. J., and M. Yogo (2019): "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 127(4), 1475 – 1515. 9

Kondor, P. (2012): "The more we know about the fundamental, the less we agree," *Review of Economic Studies*, 79(3), 1175–1207. 7

Kyle, A., and J. Lee (2017): "Toward a Fully Continuous Exchange," SSRN Working Paper. 6

Kyle, A. S. (1989): "Informed Speculation with Imperfect Competition," *Review of Economic Studies*, 56(3), 317–355. 6

Lettau, M., and S. Ludvigson (2001): "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," *Journal of Political Economy*, 109(6), 1238–1287. 20, 22, 31

Mammen, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21(1), 255–285. 21

Mondria, J. (2010): "Portfolio choice, attention allocation, and price comovement," *Journal of Economic Theory*, 145, 1837–1864. 7

Morris, S., and H. S. Shin (2002): "Social value of public information," *The American Economic Review*, 92(5), 1521–1534. 6

Ozdenoren, E., and K. Yuan (2008): "Feedback Effects and Asset Prices," *The Journal of Finance*, 63(4), 1939–1975. 7

Peress, J. (2004): "Wealth, information acquisition and portfolio choice," *The Review of Financial Studies*, 17(3), 879–914. 7, 11, 49, 50

Samuelson, P. A. (1970): "The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments," *The Review of Economic Studies*, 37(4), 537–542. 48, 49

Savov, A. (2014): "The price of skill: Performance evaluation by households," *Journal of Financial Economics*, 112(2), 213–231. 6

Sockin, M. (2015): "Not So Great Expectations: A Model of Growth and Informational Frictions," Princeton Working Paper. 7

Veldkamp, L. (2011): *Information choice in macroeconomics and finance*. Princeton University Press. 12

Yin, C. (2016): "The Optimal Size of Hedge Funds: Conflict between Investors and Fund Managers," *The Journal of Finance*, 71(4), 1857–1894. 24

Appendix

A Model Solution

Portfolio Choice We conjecture an equilibrium price which is linear in the aggregate shocks,

$$p_t = A_t + B(d_t - \mu) + C_t y_{t+1} + D_t x_{t+1} + F_t z_{t+1} \tag{15}$$

Assuming price of the form given in Equation (15), the investor derives an unbiased signal η_{pt} of y_{t+1} from the price as,

$$\eta_{pt} \equiv C_t^{-1}\left(p_t - A_t - B(d_t - \mu)\right) = y_{t+1} + C_t^{-1} D_t x_{t+1} + C_t^{-1} F_t z_{t+1}$$

This price signal has the conditional variance,

$$V\left(\eta_{pt} \mid \mathcal{I}_{it}\right) \equiv \Sigma_{pt} = C_t^{-1} D_t \Sigma_x D_t' C_t^{-1\prime} + C_t^{-1} F_t \Sigma_z F_t' C_t^{-1\prime}$$

Note that the variance of this price signal is a fixed quantity (since the coefficients are artifacts of the model, known ex ante to all investors). Given the information set $\mathcal{I}_{it}$, the investors update their beliefs of the dividend innovation y_{t+1} as per Bayesian updating to get,

$$\begin{aligned}
\mathbb{E}\left[y_{t+1} \mid \mathcal{I}_{it}\right] \equiv \mu_{it} &= \Sigma_{it}\left(\Sigma_d^{-1} \times 0 + \Sigma_{pt}^{-1}\eta_{pt} + (\zeta_{it}^2\Sigma_z + K_{it}^{-1})^{-1} s_{it}\right) \\
&= \Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1} s_{it}\right) \\
\mathbb{V}\left[y_{t+1} \mid \mathcal{I}_{it}\right] \equiv \Sigma_{it} &= \left\{\Sigma_d^{-1} + \Sigma_{pt}^{-1} + (\zeta_{it}^2\Sigma_z + K_{it}^{-1})^{-1}\right\}^{-1}
\end{aligned}$$

Further, we can express the gross payout at the end of period $t+1$ as,

$$\begin{aligned}
p_{t+1} + d_{t+1} &= A_{t+1} + B(d_{t+1} - \mu) + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2} + d_{t+1} \\
&= A_{t+1} + \mu + (B + I)\left(d_{t+1} - \mu\right) + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2} \\
&= A_{t+1} + \mu + (B + I)\left[G(d_t - \mu) + y_{t+1}\right] + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2}
\end{aligned}$$

Hence, the conditional moments of the gross payout can be expressed as,

$$\begin{aligned}
\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] &= A_{t+1} + \mu + (B + I)G(d_t - \mu) + (B + I)\mu_{it} \\
\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] &= (B + I)\Sigma_{it}(B + I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'
\end{aligned}$$

We first note that the shocks y_{t+2}, x_{t+2} and z_{t+2} do not contribute towards the conditional expectation, but are driving the conditional variance of the gross payout. On the other hand, investors form imprecise estimate for the end-of-period shock y_{t+1}, resulting in a contribution in both the

conditional moments.

In the perfect competition equilibrium (as per Lemma 1), investor i selects the optimal portfolio q_{it} given by the first order condition

$$q_{it} = \frac{1}{\rho_i}\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right]^{-1}\left\{\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] - rp_t\right\}.$$

Hence, the optimal portfolio is given as,

$$q_{it} = \frac{1}{\rho_i}\left\{(B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right\}^{-1} \times \left[\underbrace{A_{t+1} + \mu + (B+I)G(d_t - \mu) - rp_t}_{\star} + \underbrace{(B+I)\mu_{it}}_{\dagger}\right] \tag{16}$$

Market Clearing We now impose market clearing, $\int_i q_{it}di = \bar{x} + x_{t+1}$. First, note that the terms marked by $\star$ in Equation (16) are constants for the integration. Hence, we define the factor multiplying these terms – the risk tolerance weighted average precision of the gross payout,

$$\Omega_t \equiv \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di$$

We next simplify the remaining term marked by $\dagger$ in the integration in Equation (16) as,

$$\begin{aligned}
&\int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\mu_{it}di \\
&= \int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}s_{it}\right)di \\
&= \left\{\int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}di\right\}\Sigma_{pt}^{-1}\eta_{pt} \\
&+ \int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}\left(y_{t+1} + \zeta_{it}z_{t+1} + \xi_{it}\right)di \\
&= \Gamma_t\Sigma_{pt}^{-1}\eta_{pt} + \Phi_t y_{t+1} + \Psi_t z_{t+1}
\end{aligned}$$

Here, we used the fact that ξ_{it} is distributed independently of all other variables with mean zero, and defined the additional covariance terms Γ_t, Φ_t and Ψ_t (with Ω_t duplicated for reference) as,

$$\begin{aligned}
\Omega_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di \\
\Gamma_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \underbrace{(B+I)\Sigma_{it}} di \\
\Phi_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it} \underbrace{\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}} di \\
\Psi_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it} \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1} \underbrace{\zeta_{it}} di
\end{aligned}$$

As noted before, Ω_t is the risk tolerance weighted average precision of the gross payout. The terms highlighted with $\underbrace{\quad}$ indicate the additional terms in each subsequent covariance term. First, Γ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1}. Similarly, Φ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1} and the signal precision $\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}$. Lastly, Ψ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1}, the signal precision and the exposure to the public signal ζ_{it}.

We can now substitute the covariance terms Ω_t, Γ_t, Φ_t, Ψ_t and the price signal $\eta_{pt} = C_t^{-1}(p_t - A_t - B(d_t - \mu))$ in the market clearing equation to get,

$$\begin{aligned}
\bar{x} + x_{t+1} &= \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}\left(p_t - A_t - B\left(d_t - \mu\right)\right) + \Phi_t y_{t+1} + \Psi_t z_{t+1} \\
&\quad + \Omega_t\left[A_{t+1} + \mu + (B+I)G(d_t - \mu) - rp_t\right] \\
\implies \left(\Gamma_t\Sigma_{pt}^{-1}C_t^{-1} - r\Omega_t\right)p_t &= \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}A_t + \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}B\left(d_t - \mu\right) \\
&\quad - \Omega_t A_{t+1} - \Omega_t\mu - \Omega_t(B+I)G(d_t - \mu) \\
&\quad - \Phi_t y_{t+1} - \Psi_t z_{t+1} + \bar{x} + x_{t+1}
\end{aligned}$$

Let $M_t = \Gamma_t\Sigma_{pt}^{-1}C_t^{-1} - r\Omega_t$. Using the linear conjecture for the price p_t, we match coefficients as follows:

- A_t to all the constant terms: $A_t = M_t^{-1}\left[\Gamma_t\Sigma_{pt}^{-1}C_t^{-1}A_t - \Omega_t A_{t+1} - \Omega_t\mu + \bar{x}\right]$
- B to all terms with $d_t - \mu$: $B = M_t^{-1}\left[\Gamma_t\Sigma_{pt}^{-1}C_t^{-1}B - \Omega_t(B+I)G\right]$
- C_t to all terms with y_{t+1}: $C_t = -M_t^{-1}\Phi_t$
- D_t to all terms with x_{t+1}: $D_t = M_t^{-1}$
- F_t to all terms with z_{t+1}: $F_t = -M_t^{-1}\Psi_t$

Solving this yields,

$$\begin{cases} A_t = \frac{1}{r}\left\{A_{t+1} + \mu - \Omega_t^{-1}\bar{x}\right\} \\ B = (r-G)^{-1}G \\ C_t = -M_t^{-1}\Phi_t \\ D_t = M_t^{-1} \\ F_t = -M_t^{-1}\Psi_t \end{cases} \tag{17}$$

B Proofs

In order to prove Lemma 1, we first state and prove an interim utility result.

Lemma 3. *In a perfectly competitive market ($n \to \infty$), investor expected utility at date t, conditional on all date-t data is*

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tag{18}$$

Proof of Lemma 3.

From Equation (1) and Equation (10), end-of-period consumption for an investor can be represented as

$$c_{it+1} = r(\overline{w}_{it} - q_{it}'\theta_i p_t) + q_{it}'\theta_i(p_{t+1} + d_{t+1}) = r\overline{w}_{it} + q_{it}'\Pi_{it}.$$

The ex ante utility of the investor is,

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_t^-\right] = \mathbb{E}\left[\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] \mid \mathcal{I}_t^-\right]$$

That is, we calculate the ex ante utility from the interim utility using the law of iterated expectations. From Equation (3), the interim utility is given as

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i\mathbb{E}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2}\mathbb{V}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right].$$

The first order condition for optimal portfolio choice implies $q_{it} = \rho_i^{-1}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$. The first term of the interim utility is

$$\rho_i\mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{19}$$

The second term of the interim utility can be written as

$$\frac{\rho_i^2}{2}\mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' V\left(\Pi_{it} \mid \mathcal{I}_{it}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{20}$$

Taking the difference of the first term and the second term yields the result in Lemma 3

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' V\left(\Pi_{it} \mid \mathcal{I}_{it}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{21}$$

□

Proof of Lemma 1. Expand the expression for profit Π_{it} as,

$$\begin{aligned}
\Pi_{it} &= \theta_i\left[p_{t+1} + d_{t+1} - rp_t\right] \\
&= \theta_i\left[A_{t+1} + B(d_{t+1} - \mu) + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} + (d_{t+1} - \mu) + \mu - rp_t\right] \\
&= \theta_i\left[A_{t+1} + \mu + (B+I)\left[G(d_t - \mu) + y_{t+1}\right] + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} - rp_t\right] \\
&= \theta_i\left[A_{t+1} + \mu + (B+I)G(d_t - \mu) + (B+I)y_{t+1} + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} - rp_t\right]
\end{aligned}$$

The interim variance of the profit is given as,

$$\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] = \theta_i\left[(B+I)\Sigma_t(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right]\theta_i' \tag{22}$$

Here, we use the posterior variance of the dividend innovation $\Sigma_t = \mathbb{V}\left[y_{t+1} \mid \mathcal{I}_{it}\right]$. Further, it is clear from Equation (22) that the interim variance of consumption $\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ is a known quantity – it is only a function of ζ_{it} and K_{it} (in our case, ζ and K), and not a function of information revealed at the interim stage p_t or s_{it}. That is, it is a function only of the model primitives and the information set $\mathcal{I}_0$.

Next, in the expression for the conditional expected utility from Lemma 3, we decompose the conditional expected profit (3) into an expected $\mathbb{E}\left[\Pi_{it}\right]$ and a surprise component $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]$,

$$\begin{aligned}
\mathbb{E}\left[U(c_{it+1})\right] &= \mathbb{E}\left[\mathbb{E}\left[U(c_{it+1} \mid \mathcal{I}_{it})\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\{\mathbb{E}\left[\Pi_{it}\right]' + \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' - \mathbb{E}\left[\Pi_{it}\right]'\right)\right\}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left\{\mathbb{E}\left[\Pi_{it}\right] + \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right\}\right] \\
&\quad + r\overline{w}_{it}\rho_i \\
&= \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \underbrace{\mathbb{E}\left[\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right]}_{=0} \\
&\quad + \frac{1}{2}\mathbb{E}\left[\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right] + r\overline{w}_{it}\rho_i
\end{aligned} \tag{23}$$

We are interested in the second term of the ex ante expected utility in Equation (23). We will use the fact that the mean of a random variable with the central chi-square distribution is the trace

of the covariance matrix of the underlying normal variable,

$$\mathbb{E}\left[U(c_{it+1})\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2}\mathrm{Tr}\left[\mathbb{V}\left[\Upsilon_{it}\right]\right] + r\overline{w}_{it}\rho_i \tag{24}$$

$$\text{where, } \Upsilon_{it} = \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-\frac{1}{2}} \tag{25}$$

We can express $\mathbb{V}\left[\Upsilon_{it}\right]$ as,

$$\begin{aligned}\mathbb{V}\left[\Upsilon_{it}\right] &= \mathbb{V}\left[\left\{\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right\}'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-\frac{1}{2}}\right] \\ &= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\end{aligned}$$

Hence, the term of interest is the prior variance of the ex ante stochastic quantity $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$, since the prior expectation of this quantity $\mathbb{E}\left[\Pi_{it}\right]$ is a known variable ex ante. Hence, we can use the law of total variance, which says that the prior variance of the posterior expectation $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ is equal to the prior variance minus the posterior variance for Π_{it},

$$\begin{aligned}\mathbb{V}\left[\Upsilon_{it}\right] &= \left\{\mathbb{V}\left[\Pi_{it}\right] - \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right]\right\}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} \\ &= \mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\end{aligned}$$

Hence, we can express the ex ante expected utility as,

$$\mathbb{E}\left[U(c_{it+1})\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2}\mathrm{Tr}\left[\mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\overline{w}_{it}\rho_i$$

□

Proof of Lemma 2. Differentiating expected interim utility, when price p_t depends on investor i's demand yields a first order condition,

$$\begin{aligned}q_{it} &= \left[\rho_i\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] + \frac{dp}{dq_i}\right]^{-1}\left\{\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] - rp_t\right\} \\ &= \left(\rho_i\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{dp}{dq_i}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right].\end{aligned} \tag{26}$$

The term dp/dq_i, often referred to as "Kyle Lambda" is the measure of how much effect investor i's demand has on the market price of an asset.

Interim utility still takes the form

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i\mathbb{E}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2}\mathbb{V}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right].$$

However, substituting in the new expression for q_{it} from Equation (26), the first term of the interim

utility is now

$$\rho_i \mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_t \mid \mathcal{I}_{it}\right]' \left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1} \mathbb{E}\left[\Pi_t \mid \mathcal{I}_{it}\right]$$

The second term of the interim utility can be written as

$$\begin{aligned}\frac{\rho_i^2}{2}\mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right] &= \frac{\rho_i^2}{2} q_{it}' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] q_{it} \\ &= \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1} \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1} \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\end{aligned}$$

Let $\tilde{V}_i \coloneqq \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}$. Note that all terms in $\tilde{V}_i$ are known ex ante to investor i. Taking the difference of the first term and the second term yields interim expected utility

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \tilde{V}_i^{-1} \left(I - \frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tilde{V}_i^{-1}\right) \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tag{27}$$

To compute ex-ante utility, we follow the same steps as in the proof for Lemma 1. The solution is also similar, except that we replace $\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ with $\hat{V}_i \coloneqq \tilde{V}_i \left(I - \frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tilde{V}_i^{-1}\right)^{-1}$ in Equation (24) and in Equation (25). Similar to $\tilde{V}_i$, all terms in $\hat{V}_i$ are known to investor i ex ante. In this case,

$$\begin{aligned}\mathbb{V}\left[\Upsilon_{it}\right] &= \mathbb{V}\left[\left\{\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right\}' \hat{V}_i^{-\frac{1}{2}}\right] \\ &= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right] \hat{V}_i^{-1}\end{aligned}$$

Applying the law of total variance,

$$\mathbb{V}\left[\Upsilon_{it}\right] = \left(\mathbb{V}\left[\Pi_{it}\right] - \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right) \hat{V}_i^{-1}.$$

Substituting $\hat{V}_i$ for $\frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ in Equation (24) and using the new expression for $\mathbb{V}\left[\Upsilon_{it}\right]$ yields

$$\tilde{U}(\mathcal{I}_{it}) = \mathbb{E}\left[\Pi_{it}\right]' \hat{V}_i^{-1} \mathbb{E}\left[\Pi_{it}\right] + \mathrm{Tr}\left[\left(\mathbb{V}\left[\Pi_{it}\right] - \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right) \hat{V}_i^{-1}\right] + r\rho_i\bar{w}_{it}. \tag{28}$$

□

C Approximating a General Concave Utility Function

In our baseline setting, we assume that investors have mean-variance preferences over their end-of-period wealth, which itself is a function of an optimal share allocation to a risky asset.

We now show the conditions under which these mean-variance preferences represent the local approximation to a general concave utility function. The framework is that of Samuelson (1970),

which starts with a risky asset with gross payoff $\mathcal{X}$, with the pdf $f(\cdot)$. Investors in this framework have a general concave utility function $U(\cdot)$ and solve an optimal wealth share allocation problem. They assign a fraction α of their wealth to the risky asset, with the remaining being allocated to the risk-free asset. Specifically, for the situation where the first-order approximation for the mean of the risky payoff is $E(\mathcal{X}) \approx m + az$ for some scaling parameter z, define the *standardized return* $\mathcal{Z} = \mathcal{X} - m$. Then, $\mathcal{X}$ is called compact if

$$\lim_{z \to 0} \frac{E(\mathcal{Z})}{E(\mathcal{Z}^2)} = \frac{\mathcal{A}}{\mathcal{B}}, \tag{29}$$

where $\mathcal{A}$ and $\mathcal{B}$ are independent of z; and

$$\lim_{z \to 0} \frac{E(\mathcal{Z}^r)}{E(\mathcal{Z}^2)} = \sqrt{z}^{r-2} \mathcal{C}_r, \quad r = 3, 4, \ldots \tag{30}$$

where $\mathcal{C}_r$ is independent of z.

Further, in the Samuelson (1970) framework, if the payoff $\mathcal{X}$ is compact, then the optimal share $\alpha(z)$ which is the solution in the exact economy converges (as $z \to 0$) to the solution of the quadratic problem

$$\max_{\alpha} \int_0^{\infty} \left[U(m) + U'(m)(\alpha\mathcal{X} - m + 1 - \alpha) + \frac{1}{2} U''(m)(\alpha\mathcal{X} - m + 1 - \alpha)^2 \right] f(\mathcal{X}) d\mathcal{X}. \tag{31}$$

Thus, to show that the mean-variance preferences in our baseline setting (which are equivalent to solving the quadratic problem in (31)) are a valid (local) approximation to more general concave preferences, we need to define the payoff $\mathcal{X}$ and show that it is compact.

We borrow the small shock approximation of Peress (2004) to establish the conditions undre which the payoffs in our baseline model are compact.

Our environment consists the profit $\Pi \sim \mathcal{N}(\mathbb{E}[\Pi], \mathbb{V}[\Pi])$. For brevity, we ignore the index i denoting individual investor and t for time. Given the price vector p, define the inverse price matrix as

$$P^{-1} \coloneqq \operatorname{diag}(\theta p)^{-1}. \tag{32}$$

Define the quantity $\mathcal{X}$ in this economy such that the (log) return accrued to the investors when perturbed by the small deviation z (for both the first and second moments) is given as

$$\log \mathcal{X} \sim \mathcal{N}(P^{-1}\mathbb{E}[\Pi] z, P^{-1}\mathbb{V}[\Pi] P^{-1} z). \tag{33}$$

Theorem 1. *The return $\mathcal{X}$ is compact.*

The return $\mathcal{X}$ has the same distribution as in Peress (2004) and Appendix E therein proves that this return indeed satisfies the two compactness conditions (29) and (30). Since $\mathcal{X}$ is compact in our envirionment, for small values of shocks z, the investors' general preferences can be approximated

by mean–variance preferences. Within this approximation, the absolute risk aversion coefficient ρ represents the local curvature of their utility function, which can be arbitrarily dependent on their mean wealth, and the mean value of risky asset shocks.

Next, we adapt the proof of Peress (2004) Theorem (1) to our setting. We first have the expectation of $\mathcal{X}$,

$$\mathbb{E}\left[\mathcal{X}\right] = \exp\left(P^{-1}\mathbb{E}\left[\Pi\right]z + \frac{1}{2}P^{-1}\mathbb{V}\left[\Pi\right]P^{-1}z\right), \tag{34}$$

and

$$\lim_{z\to 0}\mathbb{E}\left[\mathcal{X}\right] = 1. \tag{35}$$

The standardized payoff in this economy is then given as

$$\mathcal{Z} = \mathcal{X} - 1. \tag{36}$$

For brevity, we write $\mu \equiv P^{-1}\mathbb{E}\left[\Pi\right]$ and $\sigma^2 \equiv P^{-1}\mathbb{V}\left[\Pi\right]P^{-1}$ so that $\mathbb{E}\left[X\right] = \exp\big((\mu + \frac{1}{2}\sigma^2)z\big)$. We now use the Taylor expansion for small values z to get

$$\begin{aligned}\mathbb{E}\left[\mathcal{X}^j\right] &= \exp\left(j\mu z + \frac{1}{2}\sigma^2 z j^2\right)\\ &= \sum_{i=0}^{\infty}\frac{1}{i!}\left(\mu j + \frac{1}{2}\sigma^2 j^2\right)^i z^i.\end{aligned} \tag{37}$$

Similarly, we use the Binomial Thereom to see,

$$\begin{aligned}\mathbb{E}\left[\mathcal{Z}^r\right] &= \mathbb{E}\left[(\mathcal{X}-1)^r\right]\\ &= \sum_{j=0}^{r} C_r^j(-1)^{r-j}\mathbb{E}\left[\mathcal{X}^j\right]\\ &= \sum_{j=0}^{r} C_r^j(-1)^{r-j}\sum_{i=0}^{\infty}\frac{1}{i!}\left(\mu j + \frac{1}{2}\sigma^2 j^2\right)^i z^i\\ &= \sum_{i=0}^{\infty} z^i \frac{1}{i!}\underbrace{\sum_{j=0}^{r} C_r^j(-1)^{r-j}\left(\mu j + \frac{1}{2}\sigma^2 j^2\right)^i}_{\equiv M(r,i)}.\end{aligned} \tag{38}$$

Lemma 4. *The dominant term in z for the expansion* (38) *for $r = 1$ and $r = 2$ are in the same order.*

Proof. Note that $M(r,0) = \sum_{j=0}^{r}(-1)^{r-j}C_r^j = (1-1)^r = 0$. Thus, the constant in the expansion always vanishes, for any value of $r \geq 1$.

Next, note that for $r = 1$, we have $M(1,1) = \big(\mu + \frac{1}{2}\sigma^2\big)$. As long as $\mu \neq -\frac{1}{2}\sigma^2$ we will have $M(1,1) \neq 0$, and the dominant term in z in $\mathbb{E}\left[\mathcal{Z}\right]$ is in order of z.

Similarly,

$$M(2,1) = \sum_{j=1}^{2} C_2^j (-1)^{2-j} \left(\mu j + \frac{1}{2}\sigma^2 j^2\right)$$
$$= -2\left(\mu + \frac{1}{2}\sigma^2\right) + \left(2\mu + \frac{4}{2}\sigma^2\right) \tag{39}$$
$$= \sigma^2. \tag{40}$$

As long as $\sigma \neq 0$ we will have $M(2,1) \neq 0$, and the dominant term in $\mathbb{E}[\mathcal{Z}]$ is in order of z. □

Lemma 5. *The dominant term in z for the expansion (38) for $r \geq 3$ is $\frac{r}{2} - 1$ order greater than the dominant term in $r = 2$.*

Proof. From Lemma (4), we know that the dominant term for $r = 2$ is of the order of z. For $r \geq 3$, we first use the Binomial Theorem to write

$$M(r,i) = \sum_{j=0}^{r} C_r^j (-1)^{r-j} \left(\mu j + \frac{1}{2}\sigma^2 j^2\right)^i$$
$$= \sum_{j=0}^{r} C_r^j (-1)^{r-j} \sum_{k=0}^{i} C_k^i \mu^k \left(\frac{1}{2}\sigma^2\right)^{i-k} j^{2i-k}$$
$$= \sum_{k=0}^{i} C_k^i \mu^k \left(\frac{1}{2}\sigma^2\right)^{i-k} \sum_{j=0}^{r} C_j^r (-1)^{r-j} j^{2i-k}. \tag{41}$$

Defining

$$N(r,j,l) \equiv \sum_{j=0}^{r} C_j^r (-1)^{r-j} j^l, \tag{42}$$

for any integer l, we can establish that $N(r,j,l) = 0$ for all $l \leq r-1$. To see this, consider the expression

$$(a-1)^r = \sum_{j=0}^{r} C_s^r (-1)^{r-j} a^j, \tag{43}$$

and differentiate it $l \leq r-1$ times to get

$$\sum_{j=0}^{r-l} C_j^r (-1)^{r-j} a^{j-l} \frac{j!}{(j-l)!}. \tag{44}$$

Setting $a = 1$, we get

$$\sum_{j=0}^{r-l} C_j^r (-1)^{r-j} \frac{j!}{(j-l)!} = 0. \tag{45}$$

We now prove by induction over l that $N(r,j,l)=0$. First, notice that $l=1$ results in

$$N(r,j,1)=\sum_{j=0}^{r} C_j^r(-1)^{r-j}j, \tag{46}$$

which is the same as the expression in Equation (45) with $l=1$. Next, assuming we have shown $N(r,j,s)=0$ for all $s<l$, we write the expression in Equation (45) for l as

$$\sum_{j=0}^{r} C_j^r(-1)^{r-j}\left(j^l+\sum_{s<l}a_s j^s\right), \tag{47}$$

for some coefficients a_s. Since $N(r,j,s)=0 \forall s<l$, we get $N(r,j,l)=0$. Thus, $N(r,j,l)=0$ for all $l\leq r-1$.

Hence, the first non-zero term in the expansion $M(r,i)$ will arise from i such that $2i-k>r-1$, or $i>\frac{r-1}{2}$ (since $k\in\{0,\ldots,i\}$). For odd r this will be $\frac{r+1}{2}$, and for even r this will be $\frac{r}{2}$. Thus, the dominant term in $\mathbb{E}\left[\mathcal{Z}^r\right]$ for $r\geq 3$ is in order $\frac{r}{2}$, which is $\frac{r}{2}-1$ the order of the dominant term for $r=2$. □

By Lemmata (4) and (5), we see that $\mathcal{X}$ is compact, thus proving Theorem (1).

D Skewed Payoffs

Let g be a function that maps consumption c^n under a model with normal variables into the consumption c^s under the skewed variables. As long as the new distribution of skewed model c^s is absolutely continuous with respect to normal variable model c^n, there exists a change of measure function like this. This is not saying that c^n is normal. But it is the stochastic consumption that arises out of our model with normal shocks.

Then, we can write utility under skewed distributed c^s as $U(c^s)$, which by definition of g is the same as $U(g(c^n))$. Now, define $U^s \coloneqq U(g(\cdot))$. Derive mean-variance expected utility, as in Appendix C.

The second derivative divided by the first derivative of U^s, which we will call skew-adjusted risk aversion $\hat{\rho}\coloneqq U^{s\prime\prime}/U^{s\prime}$, is not absolute risk aversion because U^s is not really preference. It is preference, convoluted with a change of measure function. So, use the chain rule to determine how adjusted risk aversion and actual risk aversion relate: $U^{s\prime}=U'g'$. Applying the chain rule a second time: $U^{s\prime\prime}=U''(g')^2+u'g''$. Now, we can write adjusted risk aversion as

$$\hat{\rho}=\frac{U^{s\prime\prime}}{U^{s\prime}}=\frac{U''(g')^2+u'g''}{U'g'}=\rho g'+\frac{g''}{g'}$$

If we use this adjusted risk aversion, we can compute expected utility with the same approximation as before, as if payoffs are normally distributed.

E Unconditional Utility in terms of Excess Returns

In this Appendix, we impose a key approximation which allows us to express the unconditional utility of the investor in terms of moments of excess, as defined in Equation (10), as opposed to profits. Recall from Equation (10)

$$R_{it} = \Pi_{it} \oslash \theta_i p_t. \tag{48}$$

Noting that p_t is known at the interim stage (in the information set $\mathcal{I}_{it}$), we start by writing the expressions for the conditional moments of R_{it}

$$\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right] = \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t, \text{and} \tag{49}$$

$$\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] = \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t p_t' \theta_i'. \tag{50}$$

We assume that the *ex-ante* variation in $\theta_i p_t$ is small relative to the other terms in the expected utility expression. Formally, this amounts to assuming that $\theta_i p_t$ is a constant from an *ex-ante* perspective. This allows us to use the law of iterated expectations and express the ex ante expectation of excess return R_{it} as

$$\begin{aligned}\mathbb{E}\left[R_{it}\right]_j &= \mathbb{E}\left[\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right]\right]_j = \mathbb{E}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t\right]_j = \mathbb{E}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]_j / \left(\theta_i p_t\right)_j\right] \\ &\approx \frac{\mathbb{E}\left[\Pi_{it}\right]_j}{\left(\theta_i p_t\right)_j}.\end{aligned} \tag{51}$$

Or equivalently,

$$\mathbb{E}\left[R_{it}\right] = \mathbb{E}\left[\Pi_{it}\right] \odot \left(\theta_i p_t\right)^{\circ(-1)}, \tag{52}$$

where $\odot$ is the Hadamard (element-wise) product of two matrices and $W^{\circ(-1)}$ represents the Hadamard (element-wise) inverse of a matrix W. Further, we use the law of total variance to express the unconditional variance of R_{it} as

$$\begin{aligned}\mathbb{V}\left[R_{it}\right] &= \mathbb{V}\left[\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right]\right] + \mathbb{E}\left[\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\right] \\ &= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t\right] + \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t p_t' \theta_i'\right] \\ &\approx \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right] \oslash \theta_i p_t p_t' \theta_i' + \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right] \oslash \theta_i p_t p_t' \theta_i' \\ &= \mathbb{V}\left[\Pi_{it}\right] \oslash \theta_i p_t p_t' \theta_i'\end{aligned} \tag{53}$$

Perfectly Competitive Markets We can now use Equations (50), (52) and (53) to express the unconditional expected utility from Lemma 1 in terms of R_{it}. We get the expression for the ex

ante expected utility in terms of excess returns as

$$\begin{aligned}\tilde{U}(\mathcal{I}_{it}) &= \frac{1}{2}\left\{\mathbb{E}\left[\Pi_{it}\right]'\mathbb{E}\left[\mathbb{V}\left[\Pi_{it}\mid\mathcal{I}_{it}\right]^{-1}\right]\mathbb{E}\left[\Pi_{it}\right]\right\}+\frac{1}{2}\operatorname{Tr}\left[\mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it}\mid\mathcal{I}_{it}\right]^{-1}-I\right]+r\overline{w}_{it}\rho_i \\ &\approx \frac{1}{2}\left\{\mathbb{E}\left[R_{it}\right]'\mathbb{E}\left[\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]^{-1}\right]\mathbb{E}\left[R_{it}\right]\right\}+\frac{1}{2}\operatorname{Tr}\left[\mathbb{V}\left[R_{it}\right]\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]^{-1}-I\right]+r\overline{w}_{it}\rho_i \quad (54)\end{aligned}$$

Imperfectly Competitive Markets Using Equation (50), we can express the modified variance $\tilde{V}_i$ as

$$\begin{aligned}\tilde{V}_i &= \mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]\odot\theta_i p_t p_t'\theta_i'+\frac{1}{\rho_i}\frac{dp}{dq_i} \\ &= \left(\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]+\frac{1}{\rho_i}\frac{dp}{dq_i}\oslash\theta_i p_t p_t'\theta_i'\right)\odot\theta_i p_t p_t'\theta_i' \\ &= \tilde{\tilde{V}}_{it}\odot\theta_i p_t p_t'\theta_i', \quad (55)\end{aligned}$$

where

$$\tilde{\tilde{V}}_{it} \coloneqq \left(\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]+\frac{1}{\rho_i}\frac{dp}{dq_i}\oslash\theta_i p_t p_t'\theta_i'\right). \quad (56)$$

Similarly, we restate $\hat{V}_i$ as

$$\hat{V}_i = \tilde{\tilde{V}}_{it}\left(I-\frac{1}{2}\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]\tilde{\tilde{V}}_{it}^{-1}\right)^{-1}\odot\theta_i p_t p_t'\theta_i' = \hat{\tilde{V}}_{it}\odot\theta_i p_t p_t'\theta_i', \quad (57)$$

where

$$\hat{\tilde{V}}_{it} \coloneqq \tilde{\tilde{V}}_{it}\left(I-\frac{1}{2}\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]\tilde{\tilde{V}}_{it}^{-1}\right)^{-1}. \quad (58)$$

We can now use Equations (50), (52), (53), (56) and (58) to express the unconditional expected utility from Lemma 2 in terms of R_{it}. We get the expression for the ex ante expected utility in terms of excess returns as

$$\tilde{U}(\mathcal{I}_{it}) \approx \mathbb{E}\left[R_{it}\right]'\hat{\tilde{V}}_{it}^{-1}\mathbb{E}\left[R_{it}\right]+\operatorname{Tr}\left[\left(\mathbb{V}\left[R_{it}\right]-\mathbb{V}\left[R_{it}\mid\mathcal{I}_{it}\right]\right)\hat{\tilde{V}}_{it}^{-1}\right]+r\rho_i\bar{w}_{it}. \quad (59)$$

F Valuing Order Flow Data

Consider an extension of the model where investors can observe data on sentiment shocks from H different data sources. Investors have the same preference and choose their risky asset investment q_{it} to maximize $\mathbb{E}\left[U(c_{it+1})|\mathcal{I}_{it}\right]$, taking the asset price and the actions of other investors as given, subject to the budget constraint (1). A given piece of data m from data source h is now a signal about x_{t+1}: $\eta_{iht}^{mx} = \psi_h^x x_{t+1}+\Gamma_h^x e_{it}^x$, with $e_{it}^x \overset{iid}{\sim} \mathcal{N}(0,I)$.

Information on sentiment shocks allows an investor i to extract a more precise signal about dividends from prices $s_{it}^p = y_{t+1}+C_t^{-1}D_t\left(x_{t+1}-\mathbb{E}\left[x_{t+1}\mid s_{it}^x\right]\right)$. While investors probably do not think about using order flow data to learn about fundamentals, they often trade against uniformed

order flow (sentiment). This is mathematically equivalent to using sentiment to extract clearer fundamental information from price and then trading on that fundamental information.

The solution of this model is a straightforward n-asset extension of the model with order flow information in Farboodi and Veldkamp (2017). Given an $N \times 1$ unbiased signal s_{it}^y about the dividend innovations y_{t+1} with precision matrix k_{it}^y and an $N \times 1$ unbiased signal s_{it}^x about the sentiment shocks y_{t+1} with precision matrix k_{it}^x, investors apply Bayes' law. They combine their prior, information in the sentiment-adjusted market price, and information on dividend innovation obtained from the data to form a posterior view about the $(t+1)$-period dividend d_{t+1}. The posterior precision is $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]^{-1} = \Sigma_0^{-1} + C_t^{-1} D_t \left(\Sigma_x + (k_{it}^x)^{-1}\right)^{-1} D_t' C_t^{-1\prime} + k_{it}^y$.

At each date t, the risky asset price equates demand with noise trades plus one unit of supply, as described by Equation (2). The equilibrium price is still a linear combination of past dividends d_t, the t-period dividend innovation y_{t+1}, and the sentiment shock x_{t+1}, as in Equation (2).

Ex-ante utility is still given by the ex-ante expectation of Equation (3). The precision variables k_{it}^y and k_{it}^x enter through the posterior variance $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$ and $\mathbb{V}\left[\Pi_t \mid \mathcal{I}_{it}\right]$. In the second term, k_{it}^y and k_{it}^x enter only through $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$. Thus, $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$ is a sufficient statistic for expected utility. The fact that the uncertainty about dividends is a sufficient statistic, and the formulation of Bayes' law for posterior precision (the inverse of uncertainty), implies that k_{it}^y and k_{it}^x affect utility in the same way, except that k_{it}^x is multiplied by $C_t^{-1} D_t D_t' C_t^{-1\prime}$. This ratio of price coefficients represents the squared signal-to-noise ratio in prices, where C is the price coefficient on the signal (future dividend) and D is the coefficient on noise (sentiment). The bottom line is that the value of sentiment data is exactly the same as the value of fundamental data, after adjusting for the signal-to-noise ratio in prices. That signal-to-noise adjustment is exactly what an OLS procedure does.

G Additional Results

G.1 Equilibrium Learning from Prices

In our equilibrium framework, prices aggregate, with noise, the dispersed information of market participants. Investors learn from both prices as well as their own data sources before making investment decisions. In Table 6, we quantify the effect of this equilibrium force on data values. Specifically, the table reports data valuations with and without price information from the perspective of an investor with \$250 million of wealth under the same assumptions about investment styles and price impact as in the baseline exercise. The row marked 'With Price Information' corresponds exactly to the values reported in Panel B of Table 2. The table shows that learning from prices exerts a quantitatively important effect (as high as 25% in some cases) on data valuations, pointing to the importance of equilibrium forces.

Table 6: **Price Information and Data Value.** Value of I/B/E/S earnings growth forecasts of Growth and S&P 500 portfolios. Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant and control variables (cay). In the first row, the value-weighted mean price of the corresponding portfolio (value-weighted mean price of the S&P500 portfolio for the last column) is supplied as an additional control variable, while this control variable is missing in the second row. Data variables being valued are the quarterly value-weighted means of I/B/E/S median forecasted earnings growth for Growth and S&P500 portfolios. Values are calculated with price impact assuming Kyle Lambda $\lambda = 1.5 \times 10^{-8}$. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Investment Style					
(in $ 000)	Small	Large	Growth	Value	S&P500	Multi-Asset
Value (Data | Controls + Price Info)	0.0	53.86	58.48	11.33	56.35	128.96
Value (Data | Controls)	0.0	49.63	54.4	8.98	47.67	102.5
Percent Change	–	8.51%	7.49%	26.26%	18.22%	25.82%

Table 7: **Price Impact and Data Value.** Value of I/B/E/S earnings growth forecasts of Growth and S&P 500 portfolios. Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant and control variables (cay). In the first row, the value-weighted mean price of the corresponding portfolio (value-weighted mean price of the S&P500 portfolio for the last column) is supplied as an additional control variable, while this control variable is missing in the second row. Data variables being valued are the quarterly value-weighted means of I/B/E/S median forecasted earnings growth for Growth and S&P500 portfolios. All values are annualized, and dollar values are reported in thousands of USD. Standard errors are calculated using a wild bootstrap.

	Investment Style					
(in $ 000)	Small	Large	Growth	Value	S&P500	Multi-Asset
Perfect Competition	0.0	110.9	116.42	52.27	340.22	1062.69
GK Price Impact	0.0	87.77	93.61	25.28	136.0	202.24
Baseline Price Impact	0.0	53.86	58.48	11.33	56.35	128.96

G.2 Price Impact

In Table 7, we explore the effect of using alternative estimates of price impact. It reports data value for the rich investor from the baseline exercise in Table 2 using the value of Kyle's Lambda as estimated by Gabaix and Koijen (2021), $\lambda = 5 \times 10^{-9}$, along with the corresponding values under perfect competition as well as our baseline price impact calibration $\lambda = 1.5 \times 10^{-8}$.

Table 8: **Return Moments.** Quarterly data between 1985–2019. Dependent variables in (13) and (14) are returns, in excess of a three-month treasury (CMT), for five portfolios: {Small, Large, Growth, Value, S&P500}. All specifications include a constant and control variables (cay and value-weighted mean price of the corresponding portfolio). In the last column, we control for a constant, cay and the value-weighted mean price of the S&P500 portfolio. Data variables being valued are the quarterly value-weighted means of I/B/E/S median forecasted earnings growth for Growth and S&P500 portfolios. The return moments reported are annualized and scaled (as indicated).

	Investment Style					
	Small	Large	Growth	Value	S&P500	Multi-Asset
$\mathbb{E}[R]$ $\times 100$	11.72	9.75	9.93	6.1	9.49	–
$\mathbb{V}[R]$ $\times 10,000$	568.15	889.16	941.92	291.71	260.5	–
$\mathbb{V}[R\|\text{Controls}]$ $\times 10,000$	558.9	870.92	921.61	288.48	256.37	–
$\mathbb{V}[R\|\text{Data + Controls}]$ $\times 10,000$	560.66	857.9	907.14	286.45	246.87	–
Utility Gain	0.0	0.009	0.009	0.004	0.026	0.082

G.3 Return Moments

Table 8 reports the moments underlying the data value calculations in Panel A of Table 2: specifically, the unconditional expected excess return, unconditional variance of returns, conditional variance of returns under two regimes: conditioning on only controls, and conditioning on controls and the data stream to be valued. These comprise the sufficient statistics to calculate the utility gain reported at the bottom, which can then be translated to the data values reported in Table 2.

H Learning from Prices

The ex ante expected utility is given as

$$\widetilde{U} = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_{it}\right]' \widehat{V}_i^{-1}\mathbb{E}\left[\Pi_{it}\right] + \text{Tr}\left[\left(\mathbb{V}\left[\Pi_{it}\right] - \mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]\right)\widehat{V}_i^{-1}\right] \tag{60}$$

where, $\widehat{V}_i^{-1} = \widetilde{V}_i^{-1} - \frac{1}{2}\widetilde{V}_i^{-1}\mathbb{V}\left[\Pi_i|\mathcal{I}_i\right]\widetilde{V}_i^{-1}$ and $\widetilde{V}_i = \mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}$.

In the simpler case of perfect competition, $\frac{dp}{dq_i} = 0$ and we have,

$$\begin{aligned}\widetilde{U} &= \mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \text{Tr}\left[\mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]^{-1}\right] + \text{constant} \\ &= \mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \text{Tr}\left[\mathbb{V}\left[\Pi_{it}\right]^{\frac{1}{2}}\mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]^{-1}\mathbb{V}\left[\Pi_{it}\right]^{\frac{1}{2}}\right] + \text{constant}\end{aligned} \tag{61}$$

For the single-asset case under perfect competition, we can see trivially that the ex ante expected

utility is

$$\widetilde{U} = \left(\mathbb{E}\left[\Pi_{it}\right]^2 + \mathbb{V}\left[\Pi_{it}\right]\right) \mathbb{V}\left[\Pi_{it}|\mathcal{I}_{it}\right]^{-1} + \text{constant} \tag{62}$$

Consider a simple single-asset case under perfect competition, where returns are represented as R, price information is represented as P and the data series to be valued is represented as D. Assume that we have already residualized all the three variables with the control series Z. Further, assume that the price signal has the structure

$$P = R + \varepsilon, \tag{63}$$

where $\varepsilon \perp R$ and $\varepsilon \sim \mathcal{N}(0, \sigma_P^2)$. This is without loss of generality, since we can always linearly transform the price series to take this form. Next, suppose we have

$$D = R + \eta, \tag{64}$$

where the mean-zero residual $\eta \perp R$ has variance σ_D^2. Without loss of generality, we assume η to be of the form

$$\eta = \beta\varepsilon + \xi, \tag{65}$$

such that $\xi \perp \varepsilon$ and $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$. Note that this directly implies

$$\sigma_D^2 = \beta^2\sigma_P^2 + \sigma_\xi^2. \tag{66}$$

In this linear setting, we also have $\beta = \rho_{PD}\frac{\sigma_D}{\sigma_P}$, where $\rho_{PD} = \text{Corr}(\varepsilon, \eta)$. Under this setting, we can apply Bayes' Law to see,

$$\mathbb{V}\left[R|D\right]^{-1} = \mathbb{V}\left[R\right]^{-1} + \mathbb{V}\left[D|R\right]^{-1}. \tag{67}$$

Note that $\mathbb{V}\left[D|R\right] = \sigma_D^{-1}$. The difference $\mathbb{V}\left[R|D\right]^{-1} - \mathbb{V}\left[R\right]^{-1}$ is linearly proportional to the value of the data series D, without using the information in prices P.

To get the value of data D while conditioning on P, we need to extract the unbiased signal for R from D after conditioning on P. This signal can be obtained as

$$\frac{D}{1-\beta} - \frac{\beta}{1-\beta}P = R + \frac{\xi}{(1-\beta)} \tag{68}$$

Thus, we again apply Bayes' Law to see,

$$\mathbb{V}\left[R|D,P\right]^{-1} = \mathbb{V}\left[R|P\right]^{-1} + \mathbb{V}\left[D|R,P\right]^{-1}. \tag{69}$$

Note that $\mathbb{V}\left[D|R,P\right] = \frac{\sigma_\xi^2}{(1-\beta)^2}$. The difference $\mathbb{V}\left[R|D,P\right]^{-1} - \mathbb{V}\left[R,P\right]^{-1}$ is linearly proportional to the value of the data series D, while using the information in prices P.

Thus, we have

$$\text{Value}(D) < \text{Value}(D|P)$$
$$\Leftrightarrow \rho_{PD}^2 \left(1 - \frac{\sigma_D^2}{\sigma_P^2}\right) - 2\rho_{PD}\frac{\sigma_D}{\sigma_P} \equiv \Delta > 0$$

In particular, if $\sigma_D < \sigma_P$, for all values of the correlation ρ_{PD} which are sufficiently negative, the value of data with price information is higher than the value of data without price information. On the other hand, if $\sigma_D \geq \sigma_P$, data value with price information is also higher than without price information for some positive values of ρ_{PD}.

In Figure 1, we plot the results from a simulation with the iid DGP $R \sim \mathcal{N}(0,1)$ and parameter values $\sigma_D = 0.5$ and $\sigma_P = 1$. We use $10,000$ sample points, and vary the value of ρ_{PD} between $[-0.8, 0.8]$. We plot the difference in the conditional precisions, $\mathbb{V}\left[R|D,\mathcal{I}\right]^{-1} - \mathbb{V}\left[R|\mathcal{I}\right]^{-1}$, which we use as the proxy for value of the data series $\text{Value}(D|\mathcal{I})$. Even in this simple example, we see the non-monotonic relationship between the value of D while conditioning on P, and the correlation ρ_{PD}. More importantly and counterintuitively, the value of data with price information in the conditioning set exceeds the value of data without conditioning on price information for a large range of (negative) values of ρ_{PD}.

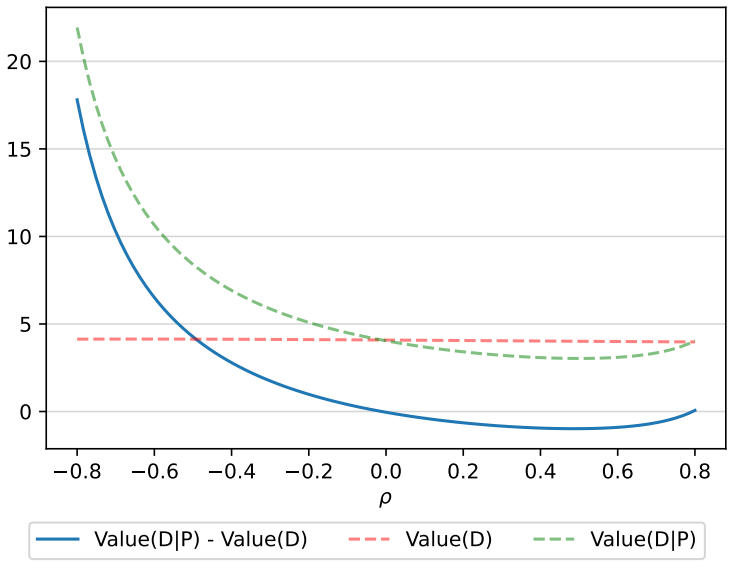

Figure 1: **Learning from Prices and Data Value:** Value of a randomly generated data series is plotted, with the price information and without. The solid line indicates the difference of the two data values.