Data, Markups, and Asset Prices

Alexandre Corhay, Kejia Hu, Jun E. Li, Jincheng Tong, and Chi-Yang Tsou*

Abstract

This paper studies the implications of data technology for firm dynamics and asset prices. We develop a heterogeneous firm model in which firms optimally hire data scientists to learn about unobserved consumer preferences. Data enhances firms' demand forecasting accuracy, enabling them to charge higher markups. Firms that are constrained in expanding production capacity have stronger incentives to hire data scientists. This results in countercyclical data scientist hiring, which amplifies firms' exposure to aggregate risk via the operating leverage channel. Using a novel dataset that tracks firms' employment of data scientists, we document three key empirical findings that support the model's main mechanisms: firms with a higher share of data scientists exhibit larger markups, higher information quality, and higher stock returns.

JEL Codes: E2, E3, G12

Keywords: Data Scientists, Markups, Firm Dynamics, Asset Prices

First Draft: September 30, 2023

This Draft: June 8, 2025

*Alexandre Corhay (alexandre.corhay@rotman.utoronto.ca), Rotman School of Management, University of Toronto; Jun E. Li (jun.li@wbs.ac.uk), Warwick Business School, University of Warwick; Kejia Hu (kejia.hu@sbs.ox.ac.uk), Saïd Business School, University of Oxford; Jincheng Tong (jincheng.tong@rotman.utoronto.ca), Rotman School of Management, University of Toronto; and Chi-Yang Tsou (chi-yang.tsou@manchester.ac.uk), Alliance Manchester Business School, University of Manchester. We thank Bing Han, Po-Hsuan Hsu, Péter Kondor, Maria Marchica, Charles Martineau, and Gill Segal for their helpful comments and seminar and conference participants at the University of Warwick, MFA meeting, Labor and Finance Conference, Academia Sinica, SUFE, University of Manchester, National Chengchi University, and National Taipei University. We also thank Yijian Sun for excellent research assistance. Jun E. Li gratefully acknowledges the Pujiang grant for financial support. Jincheng Tong gratefully acknowledges the TD MDAL grant from the Rotman School of Management and Social Sciences and Humanities Research Council of Canada for financial support. Chi-Yang Tsou gratefully acknowledges the AMBS Research Support Fund for financial support.

1 Introduction

The significance of data in the modern economy has grown dramatically in recent years. Data is crucial in enhancing companies' understanding of consumer behavior, enabling them to optimize pricing strategies. For instance, Uber employs dynamic pricing algorithms that leverage customer data and historical activity patterns. Similarly, airlines and retailers utilize consumer data and sophisticated algorithms to deliver personalized recommendations and boost sales performance. These examples illustrate a key benefit of data investment: by analyzing customer behavior, firms can extract monopolistic rents and increase their price markups. To examine the impact of data on firms' corporate policies, market power, and asset prices, we extend the standard q-theory model by allowing firms optimally hire data scientists and learn about consumer demand. We empirically test the model's predictions using a novel dataset that tracks firms' employment of data scientists.

Our model provides a framework for understanding the role of data in affecting firms' endogenous decisions. Firms can optimally choose to hire data scientists, which helps them forecast consumer demand. There's an unobservable component, consumer habit or taste, along with the standard elastic component of the demand curve.[1] This habit component is inelastic to price, and its relative importance depends on the firm's ability to accurately forecast unobserved consumer taste. When a firm better aligns its products with current consumer taste, this inelastic component plays a more important role, allowing the firm to charge higher price markups. Thus, the demand side of our model provides a theory of demand forecasting and markups. For the production technology, our model shares several common features with the production-based asset pricing literature, including capital adjustment costs and heterogeneous firms due to idiosyncratic firm-level TFP shocks.

Data plays a crucial role in resolving uncertainty about consumer taste and improves demand forecasting. In our setup, a firm observes a noisy signal of the true consumer taste. Following Farboodi and Veldkamp (2021), we assume this signal contains both the true taste and a noise component, with the signal's precision increasing with the efficient units of data generated by the firm's data scientists. This mechanism incentivizes firms to hire data scientists and accumulate data, which ultimately leads to higher markups. To emphasize the role of data scientists in learning and increasing market power, we assume they do not directly contribute to production. Consequently, there is no factor income associated with their activities.

The optimality condition for hiring data scientists introduces a novel data-Q relationship

[1] We use habit and taste interchangeably. The notion of habit in the demand curve is very similar Ravn, Schmitt-Grohé, and Uribe (2006).

analogous to the classic Q theory of investment. This relationship states that the marginal cost of hiring an additional data scientist must equal the present value of the markup increase derived from improved demand forecasts. The data-Q relation also yields predictions about the joint dynamics of productivity shocks and data scientist hiring. Specifically, firms experiencing production bottlenecks (sequence of adverse firm-level TFP shocks) have stronger incentives to increase current profits by raising price markups. To achieve this, firms hire more data scientists to enhance their data processing capabilities, improving their ability to track consumer demand and strengthen market power. Therefore, our model predicts that data scientist hiring should be relatively more countercyclical than other cyclical business quantities, such as capital investment.

The countercyclical hiring of data scientists has important implications for the risk profile of data-intensive firms. On one side, data scientists enable firms to charge higher markups and maintain profitability during economic downturns, providing a valuable hedge for investors during periods of high marginal utility. However, the countercyclical nature of data scientist hiring also introduces a labor-induced operating leverage effect that amplifies the cyclicality and risk in firms' cash flows. Our quantitative analysis reveals that the operating leverage effect dominates the hedging benefit, suggesting that firms with higher proportions of data scientists should earn greater returns in equilibrium.

Given the intangible nature of data, measuring a firm's data technology presents an empirical challenge. To address this challenge, we examine firms' workforce composition, specifically focusing on employees with data analytical skills, whom we term as data scientists. Under the assumption that the amount of data a firm can process is determined by its labor force with such skills, our focus on data scientists should provide a good proxy for a firm's unobserved data technology. We use a detailed occupation-level dataset to construct a firm-year measure of employees with data analytical skills. Our key metric is the data scientist ratio, defined as the proportion of data scientists relative to total employees.

We document several novel empirical facts regarding data scientists, firm dynamics, and returns that are consistent with our model predictions. First, firms with more data scientists exhibit higher markups and profit margins, suggesting that data accumulation indeed strengthens firms' market power. Second, we find that firms hiring more data scientists can improve their forecast accuracy. Third, we show that firms with higher data scientist ratios are riskier and earn higher expected stock returns. A long-short portfolio based on the data scientist ratio yields an annual excess return of approximately 4%, which remains significant after controlling for popular equity risk factors.

Finally, we calibrate our model to match salient features of the data regarding the cross-sectional variation in data scientist ratio, markups and other prominent firm characteristics.

We find that our model can quantitatively account for the positive association between data scientist ratio, information quality, markups, and expected stock returns. In addition to the quantitative success, we further test our model's predictions regarding the cyclicality of data scientist hiring both at the aggregate time series and cross-sectional firm level. At the aggregate level, we find that the IT-related investment to total investment ratio is countercyclical. At the firm level, we show that firms with more data scientists experience lower firm-level productivity shocks.

Related Literature

Our paper contributes to the growing literature studying the impact of data on the economy and the financial markets (see Veldkamp and Chung (2022) for a recent review). We model data as a tool that mitigates information frictions and enhances firms' learning about unobserved fundamentals, building on the framework in Farboodi and Veldkamp (2021), which emphasizes data's role in reducing uncertainty. Abis and Veldkamp (2024) also use employment information of data-related jobs to infer the value of data. Begenau, Farboodi, and Veldkamp (2018) demonstrate how data's impact on uncertainty reduction disproportionately benefits large firms. Babina, Fedyk, He, and Hodson (2024) show firms that invest more in AI increase their exposure to the systematic risk, measured by market betas. Brynjolfsson and McElheran (2019) find that data-driven decision-making enhances efficiency. Bian et al. (2024) studies the effect of data sharing on creating comovements among data-connected firms. While prior work associates data technology with gains in efficiency and productivity, our research highlights data accumulation's role in increasing market power. Additionally, we address the risks associated with data investment and its implications for asset pricing through the learning channel, filling a gap in the literature.

This paper belongs the production-based asset pricing literature based on q-theory. For instance, Zhang (2005) explains the value premium within a q-theory framework, while Ai, Li, and Tong (2022) and Kogan, Li, and Zhang (2023) address both value and profitability premiums. Our concept of data scientists aligns with studies examining labor's impact on firm dynamics and stock returns. For example, Belo, Li, Lin, and Zhao (2017) find that industries with more high-skilled workers tend to amplify the negative relationship between hiring rates and future stock returns, and Donangelo, Gourio, Kehrig, and Palacios (2019) and Kuehn, Simutin, and Wang (2017) analyze how labor share and exposure to labor market conditions, respectively, relate to equity returns. However, our study diverges from these by examining data scientists' influence on firm dynamics via market power rather than as a traditional labor input. In our model, firms' optimal data acquisition resembles investment

q-theory, with our "data-Q" relation indicating that the present value of expected markups and reductions in demand uncertainty drive variations in data investment across firms and over time.

The concept of data relates to the literature on intangible capital. For example, Eisfeldt and Papanikolaou (2013) highlight the significance of organizational capital for firm dynamics, particularly due to the risks posed by key talent departures. Crouzet and Eberly (2023) suggest that intangible assets account for significant fluctuations in the gap between investment rates and Tobin's q. Our paper introduces a distinct perspective, positioning data as a unique role that enhances firms' ability to forecast demand, which increases firms' market power. This approach expands the literature by focusing on data's specific role in improving demand prediction and boosting firm market power.

Our article relates to the growing literature in finance studying the impact of market power on asset prices in equilibrium models with production and strategic interactions, e.g., Aguerrevere (2009), Bustamante and Donangelo (2017), Loualiche et al. (2016), Corhay, Kung, and Schmid (2020), Dou and Ji (2021), Dou, Ji, and Wu (2021). In these studies, market power matters for risk because it amplifies exposure to aggregate productivity risk. Earlier works have not examined the risks associated with data investment. We provide a theory of heterogeneous exposures to aggregate risk when firms can endogenously acquire data and improve the accuracy of information. Corhay, Li, and Tong (2022) explores the impact of asset pricing on exogenous shocks to aggregate markups. We analyze the role of data investment in securing market power which leads to endogenous variation in markups and study its asset pricing implications. Finally, Eeckhout and Veldkamp (2022) develops a framework to study the interaction between data and imperfect competition. We share some of the insights emerging in their static framework: data leads to risk reduction which tends to make firms safer. In addition, our dynamic model provides a more complete characterization of the risks associated with data accumulation: countercyclical data investment constitutes a form of operating leverage that increases firms' risks and expected stock returns. Our theoretical model also provides guidance on how to isolate the two competing effects empirically.

The remainder of our paper is organized as follows. In Section 3, we present some stylized empirical facts and introduce data construction and summary statistics. We introduce our model in Section 2 and provides a quantitative analysis of the model in Section 4. Section 5 concludes.

2 The Model

We introduce a quantitative production-based asset pricing model that includes data scientists to shed light on the role of data on consumer demand and firm dynamics. Our model is grounded in the investment-based asset pricing literature, such as Zhang (2005), Ai, Li, and Tong (2022), and Kogan, Li, and Zhang (2023), and incorporates core features of these models, including convex capital adjustment costs, neoclassical production technology, and heterogeneous firms due to firm-specific productivity shocks. We contribute to this literature by incorporating new elements on the consumer demand side, examining the role of data scientists in demand forecasting, markups, and stock returns.

2.1 Production Technology

Time is infinite and discrete. The economy is populated by a continuum of firms indexed by i. A firm uses capital K_{it} to produce the product Y_{it}. The production technology follows the standard Cobb-Douglas function,

$$Y_{it} = Z_{it} A_t K_{it}^{\alpha},$$

where α controls the share of capital in the production. Firm-level productivity is denoted as Z_{it}, and A_t refers to aggregate productivity.

To focus on the effect of data scientists, we impose the assumption that firm's demand for normal labor, which refers to the labor force that performs production related tasks, is fixed and normalized to one. As a result, the number of data scientists in our model can also be interpreted as the ratio of data scientists to total employees, consistent with our empirical measure of DS ratio.

The law of motion of capital K_{it} is given by

$$K_{it+1} = (1 - \delta_K) K_{it} + I_{it}, \tag{1}$$

where δ_K is the capital depreciation rate, and I_{it} is the investment. Capital investment is subject to adjustment costs. We assume the standard quadratic form of capital adjustment cost $G^K(I_{it}, K_{it})$. The quadratic adjustment cost function ensures that the investment adjustment cost is zero in steady state when the investment rate equals the capital depreciation rate.

Data Scientists Firms maintain a pool of data scientists N_{it}. We interpret N_{it} as data scientists scaled by the total number of employees. New hires are denoted as H_{it}. δ_N is the fraction of existing data scientists that leave the firm. The law of motion of data scientists is

$$N_{it+1} = (1 - \delta_N)N_{it} + H_{it}, \tag{2}$$

Changes in the current data scientist pool are subject to convex adjustment costs $G^N(H_{it}, N_{it})$.

Data scientists produce data that will eventually improve the accuracy of demand forecasting as shown later. Data production technology can be interpreted as a combination of data collection and analysis. Following Jones and Tonetti (2020), we assume that data production is dependent on the past unit sold of the firms, Y_{it-1}. The idea that larger firms produce more data is consistent with the data production function as in Abis and Veldkamp (2024). The one period lag between quantity sold and data production is intuitive since it takes time for data scientists to analyze the volumious transaction data. For a firm i with a total number of data scientists N_{it}, the amount of data produced is given by

$$\omega_{it} = \nu_0 N_{it}^{\nu_1} Y_{it-1}^{1-\nu_1}, \tag{3}$$

where ν_0 is a scaling parameter, and $\nu_1 \in (0, 1)$ controls the share of data scientists in data production.

2.2 Data and Demand Curve

2.2.1 The Demand Curve

We focus on the firm's demand without explicitly modeling the representative consumer's consumption and saving decisions. The consumer derives utility from a continuum of differentiated goods indexed by $i \in [0, 1]$

$$\mathcal{Y}_t = \left[\int_0^1 (Y_{it} - \Theta_{it})^{1-1/\eta} di \right]^{1/(1-1/\eta)},$$

where Y_{it} refers to consumption goods i produced by firm i, η controls the demand elasticity, and Θ_{it} denotes the good-specific habit. We assume that each firm produces one differentiated good, and consumers exhibit a good-specific level of habit.[2] This good-specific habit arises endogenously as an equilibrium outcome of the firms' heterogeneous decisions. In Section 2.2.2, we discuss in detail how each firm is equipped with a forecasting technology to track the representative consumer's habit. Although the representative consumer's habit is the same across all firms in any given period, differences in the firms' efforts to track this

[2] Alternatively, one could assume that each firm produces a bundle of goods, with these bundles differentiated across firms by different levels of habit.

habit result in good-specific (or firm-specific) demand. We also discuss how firms can take actions to increase the good-specific habit, Θ_{it}, effectively making consumers more inclined to consume their product.

For any given level of $\mathcal{Y}_t$, the purchase of each good variety $i \in [0,1]$ in period t must solve the expenditure minimization problem,

$$\min_{Y_{it}} \int_0^1 P_{it} Y_{it} di - P_t \left[\int_0^1 (Y_{it} - \Theta_{it})^{1-1/\eta} di \right]^{1/(1-1/\eta)}.$$

The solution to the minimization problem yields the demand curve for consumption good Y_{it}, that is,

$$Y_{it} \leq \left(\frac{P_{it}}{P_t} \right)^{-\eta} \mathcal{Y}_t + \Theta_{it}. \tag{4}$$

A firm's product demand consists of two components: one that responds to price changes and another that remains inelastic. When the inelastic component accounts for a larger share of total demand variation, the overall demand elasticity decreases, enabling the firm to set higher price markups.[3]

2.2.2 Forecasting Technology

Our empirical findings indicate that firms with higher DS ratios tend to achieve greater precision in sales forecasting and more improvements in forecast revisions. These observations motivate us to establish a connection between data and demand forecasting.[4]

Firms choose technology θ_{it} to align closely with the unobserved taste of the representative consumer x_t. For instance, x_t can represent the popular color among clothing shoppers for the current season, while θ_{it} represents the color companies predict will be popular. If firms make accurate predictions and their chosen color closely matches the true consumer preference, they will enjoy higher sales volumes.

The good-specific habit effect Θ_{it} is modeled as the squared distance between the firm's choice of θ_{it} and the representative consumer's taste x_t:

$$\Theta_{it} = \bar{\Theta} - (\theta_{it} - x_t)^2. \tag{5}$$

[3] Our framework is very similar to the deep habit literature, as in Ravn et al. (2006). Deep habit specifications have been widely used in macroeconomics and asset pricing, such as Heyerdahl-Larsen (2014), van Binsbergen (2016), Gilchrist et al. (2017), and Crouzet and Mehrotra (2020).

[4] Our model views data as a tool to alleviate information frictions. Similar arguments can be found in Kozlowski et al. (2018), Begenau et al. (2018), Farboodi, Mihet, Philippon, and Veldkamp (2019), Farboodi and Veldkamp (2021), Veldkamp and Chung (2022), Eeckhout and Veldkamp (2022), and Abis and Veldkamp (2024).

The unobserved aggregate consumer taste x_t follows a normal distribution $x_t \sim \mathcal{N}(0, \sigma_\theta^2)$ and it is *i.i.d.* across time.[5] It will become clearer later that this can also be interpreted as a tracking problem, where firms select the tracking technology θ_{it} to minimize their tracking error with respect to consumer taste x_t.

Firms cannot observe x_t directly. However, they receive noisy signals s_{it} that are informative about x_t. For each data point $m \in [1 : \omega_{it}]$, the signal $s_{i,t,m}$ contains the true consumer taste plus a noise term $\varepsilon_{i,t,m}^s$,

$$s_{i,t,m} = x_{t+1} + \varepsilon_{i,t,m}^s,$$

where $\varepsilon_{i,t,m}^s$ is *i.i.d.* across firm and time. The noise follows a normal distribution $\varepsilon_{i,t,m}^s \sim N(0, \sigma_s^2)$.

Data can help increase the precision of the total signal s_{it}. Specifically, if firm i has ω_{it} amount of data, the corresponding distribution of the signal becomes

$$s_{it} = x_{t+1} + \varepsilon_{it}^s, \quad \varepsilon_{it}^s \sim N(0, \frac{\sigma_s^2}{\omega_{it}}).$$

Effectively, with more data points, the signal becomes more precise. To avoid the complex strategic interaction due to inferring x_t from other firms' actions, we assume that firms do not observe each other's technology decision θ_{it}. Otherwise firms may choose not to learn and just observe others decisions.

Some studies link data to technological advancements that enhance productivity or product quality (e.g., Jones and Tonetti (2020), Agrawal, McHale, and Oettl (2018)). In theory, under sticky prices, firms can achieve higher markups by using data to improve production efficiency and reduce marginal costs, such as through better inventory management. However, identifying the exact source of these efficiency gains is beyond the scope of this paper.

2.2.3 Data and Forecasting Precision

Data scientists produce data that reduces the noise of the signal s_{it} within the firm. Firm's forecast precision of demand curve can be captured by the posterior variance of x_t.

[5] Farboodi and Veldkamp (2021) assume x_t is persistent and introduce a white noise term to prevent firms from inferring the true x_t by observing past sales. In our model, we assume x_t to be i.i.d. However, it does not mean data itself is not persistent. We will show that the persistent in the accumulation of data scientists allows firms to continuously keep and gather data, which means data scientists do not forget what they have learned. Additionally, we do not need to assume the additional white noise term. This is because x_t is assumed to be i.i.d. in our model. Data will not be stored and firms will never infer the true $\Theta_{i,t}$, even if they can observe the history of all variables.

The precision of firm i's forecast of x_t is

$$\Omega_{it} = \mathbb{E}[(\mathbb{E}[\theta_t | \mathcal{I}_{it}] - x_t)^2]^{-1},$$

where Ω_{it} is the posterior variance of the x_t conditional on information set $\mathcal{I}_{it}$. The inverse of Ω_{it} is the conditional variance of the forecasting error.

The dynamics of the posterior, following the standard Bayesian updating rule, evolves as follows,

$$\Omega_{it} = \sigma_\theta^{-2} + \sigma_s^{-2} \omega_{it}, \tag{6}$$

where ω_{it} is the amount of data produced by the data scientists by firm i. Clearly, the higher the data produced ω_{it}, the more precise the demand signal is. As Equation (6) shows that firms can accumulate data by keep data scientists.

2.3 Firm's Problem

A firm optimally chooses their taste tracking technology θ_{it}, hiring decision of conventional labor L_{it} and data scientist H_{it}, and investment decisions I_{it}, to maximize the net present value of dividends,

$$V_{it} = \max_{\theta_{i,t+j}, H_{i,t+j}, I_{i,t+j}, L_{i,t+j}} \mathbb{E}_{it} \left[\sum_{j=0}^{\infty} M_{t,t+j} D_{i,t+j} \right], \tag{7}$$

subject to equation (1), (2), (4), (6), and dividend D_{it} is defined as

$$D_{it} = P_{it} Y_{it} - I_{it} - G^K(I_{it}, K_{it}) K_{it} - W_t^N N_{it} - G^N(H_{it}, N_{it}) N_{it} - F, \tag{8}$$

where W_t^N is the wage rate data scientists, F is the fixed cost, and $M_{t,t+j}$ is the stochastic discount factor between period t and $t+j$. We introduce exogenous wage processes for data scientists and workers in the calibration section.

2.4 Optimality Conditions

This section derives the conditions that characterize the optimal firms' choices. We then provide two propositions that analyze the trade-off between data investment, uncertainty reduction, and markups.

We solve the optimal tracking problem similarly to Farboodi and Veldkamp (2021), which allows us to replace the good-specific habit Θ_{it} with the precision of the information Ω_{it}.

$$Y_{it} \leq \left(\frac{P_{it}}{P_t}\right)^{-\eta} \mathcal{Y}_t + (\bar{\Theta} - \Omega_{it}^{-1}). \tag{9}$$

The detailed derivations can be found in Appendix A.1.

With data technology, firms adjust the composition of their total demand in response to economic shocks. A firm with better data technology that tracks consumers' tastes more closely has higher information quality Ω_{it} and lower posterior variance of consumer taste. This essentially increases the weight of inelastic demand and allows firms to charge higher markups.

Another way of interpreting demand function (9) is to consider a firm's customer base. A firm's product demand comes from customers who react to price changes and those who don't. When price-insensitive customers drive more demand fluctuations, overall demand becomes less elastic, letting firms charge higher markups. Data technology gives firms the tools to alter this demand mix. By tracking consumer tastes more precisely, firms with more advanced data technology can better identify and cater to price-insensitive customers, increasing this group's influence on overall demand and charging higher markups.

We formally show the impact of the data-learning channel on data investment and markups in the following propositions.

Proposition 1. *(Q-theory of Data Scientists)*

The optimal hiring rate of data scientists satisfies

$$q_{it}^N = \mathbb{E}_{it}\left[M_{t,t+1}\left\{\nu_0\nu_1\sigma_s^{-2}\lambda_{it+1}\Omega_{it+1}^{-2}N_{it+1}^{\nu_1-1}Y_{it}^{1-\nu_1} - W_{t+1}^N - \frac{\partial G^N}{\partial N_{it+1}} + q_{it+1}^N(1-\delta_N)\right\}\right], \tag{10}$$

where $q_{it}^N = \frac{\partial G^N}{\partial H_{it}}$ is the marginal q of data scientists, and λ_{it} is Lagrangian multiplier associated with the demand equation which denotes the shadow value of demand.

Proof. See Appendix A.1. □

The left-hand side is the marginal q for data scientists. The right-hand side is the present value of all future benefits of hiring an additional data scientist. The first term states that adding data scientists raises a firm's market power since better information allows firms to better forecast consumer's tastes and charge higher markups. Clearly, if markups are countercyclical, the benefits of data investment that are associated with markups should also be countercyclical, thus inducing countercyclical variations in data investment.

Expanding the pool of data scientists also enlarges the wage bill and the increase in cost also plays a role in determining a firm's hiring decision which shows up in the second term. In bad times, too much wage bill of data scientists amplifies the risks in a firm's cash flows and investors require a higher expected return to hold the firm's equity.

Proposition 2. *(Pricing Equation)*

The price charged by a firm can be determined by

$$P_{it} = \lambda_{it} - \tilde{\lambda}_{it} + MC_{it},$$

where $\tilde{\lambda}_{i,t} = \mathbb{E}_t[\nu_0(1-\nu_1)\sigma_S^{-2}\lambda_{it+1}\Omega_{it+1}^{-2}N_{it+1}^{\nu_1}Y_{it}^{-\nu_1}]$.

Proof. See Appendix A.1. □

Proposition 2 shows that the pricing equation includes an additional wedge term, $\tilde{\lambda}_{it}$, which is different from the classical Dixit-Stiglitz framework. This wedge term arises because firms rely on both data scientists and past sales to produce valuable data for forecasting demand curves. This setup introduces an intertemporal tradeoff: a firm can boost its market power today by raising prices and reducing production, but this approach could weaken future market power by reducing the volume of transaction data can be used for predicting customer demand. As a result, exploiting the market power by lower production today can reduce firm's data accumulation and its future market power. This channel is captured by the wedge term $\tilde{\lambda}_{it}$ in the pricing equation.

Proposition 3. *(Markup and Data)*

The firm-level markup is determined by the following equation

$$\mu_{it} = \left(1 + \frac{\tilde{\lambda}_{it}}{P_{it}} - \frac{1}{\eta}\frac{Y_{it}}{Y_{it} - (\bar{\Theta} - \Omega_{it}^{-1})}\right)^{-1} \tag{11}$$

Proof. See Appendix A.1. □

Proposition 3 establishes the relationship between firm-level markups and the data produced by data scientists. Firms with more data scientists can increase their precision Ω_{it}, which leads to a higher markup μ_{it}. Intuitively, when firms have more data scientists, they can predict their demand curve more accurately, which can improve their pricing strategy. As a result, they can charge a higher markup. The role of the wedge term $\tilde{\lambda}_{it}$ follows the intuition we have discussed under Proposition 2, it lowers down the firm's markup.

Additionally, Proposition 3 highlights two channels for countercyclical markups. First, Equation (11) demonstrates that firm-level markup μ_{it} is negatively related to output Y_{it}.

Second, when a firm experiences a negative idiosyncratic productivity shock Z_{it}, it optimally increases its hiring of data scientists to enhance market power. This occurs because the factor input K_{it} is predetermined, leaving the firm with limited options to increase its revenue. To maximize the present value of the firm, it focuses on the other choice, that is, the price of its goods. By hiring more data scientists, the firm can strategically increase its price markup. This intuition corroborates with Proposition 1 that the marginal q of data scientists is tightly linked to the market power.

3 Empirical Findings

The three propositions in the previous section develop a set of theoretical predictions that we test in this section. Our model predicts that data scientists can improve the precision of managerial forecasts, which is a proxy for the quality of managerial information, leading to a positive association between data scientists and markups. The data-Q equation (10) also highlights the potential implications of data scientists and risk. If markups are countercyclical, the benefits of data investment that are associated with markups should also be countercyclical, thus inducing a form of operating leverage and amplifying firms' risks. To test these predictions, we first discuss our measure of firm-level data scientists. Using this new measure, we document several new findings regarding data technology, information quality, markup and stock returns.

3.1 Data Source

Job occupation data We obtain occupation-level job data from Revelio Labs, a major provider of labor data that continuously gathers online profiles and resumes from platforms such as LinkedIn. Revelio Labs categorizes jobs into 150 distinct occupations.[6] The data becomes comprehensive since 2008 and is aggregated at the firm-occupation-month level, covering current headcounts, as well as employee exits and new hires within a firm for each month. We identify data scientists within this dataset by selecting occupations that perform roles similar to data scientists, such as data scienitsts, economists, data analysts, etc. Section C.1 of the Appendix provides a detailed list of data science-related occupations. To merge this data with Compustat, we aggregate the monthly employment data to an

[6] To enhance the data's representativeness, especially considering potential biases in online profiles, Revelio Labs employs a de-biasing technique. This method adjusts the recorded employee counts based on the likelihood of individuals having online profiles, with adjustments according to the distribution from the Bureau of Labor Statistics. Additional details on job categories and methodologies are available on Revelio Labs' website: https://www.data-dictionary.reveliolabs.com/methodology.html.

annual frequency. Our key measure, the data scientist ratio, is defined as the ratio of data scientists to total employees for each firm at the end of each calendar year.

Stock returns and firm fundamentals Our sample consists of firms that lie in the intersection of Compustat, the Center for Research in Security Prices (CRSP), and the Revelio database, dropping observations that lack necessary financial data and stock returns. We obtain accounting data from Compustat and stock price data from CRSP. Our sample firms include those with nonmissing data scientist ratios and nonmissing NAICS two-digit industry classification codes, as well as those with domestic common shares (SHRCD = 10 and 11) trading on NYSE, AMEX, or NASDAQ. Following the stream of literature of empirical asset pricing, we exclude financial firms. We require firms have at least two years of observations in Compustat.

IBES and forecasts We collect data on management's sales forecasts and actual sales from the IBES (Institutional Brokers Estimate System) Guidance database. Managers' sales forecast errors are used as a proxy for the precision of firms' demand forecasts, reflecting how accurately firms forecast consumer demand for their goods. The idea is that firms with more data scientists can provide better information to managers, improving the accuracy of these forecasts. To address the potential issue that firms with more analyst coverage may have an information advantage, we control for the number of analysts covering a firm, which is also collected from the IBES database.

Job posting data To complement our firm-level employment measures, we utilize job posting data from Lightcast to capture the content of data-scientist-related roles. Lightcast aggregates millions of online job advertisements across platforms and provides structured fields such as job title, employer, location, and most importantly, full-text job descriptions. We implement a text analysis framework to extract keywords based on these descriptions, categorizing terms into four categories: core data science, analytical methods, business topics, and technical infrastructure. We then compute the frequency of keywords in each category to reflect the data science activities demanded by firms. The Lightcast data provides evidence of what data scientists are actually recruited to do by capturing the specific tasks, tools, and objectives emphasized in job postings. This detail complements our key measure (e.g., the data scientist ratio) and helps consolidate our interpretation of the hiring of data scientists.[7]

[7] For more details on the Lightcast data, see https://lightcast.io/products/data/overview.

Macroeconomic Data Macroeconomic data are from the Federal Reserve Economic Data (FRED) maintained by the Federal Reserve in St. Louis.

3.2 Summary Statistics

Table 1, Panel A reports pooled summary statistics. Specifically, Panel A reports the pooled mean, median, standard deviation (Std), 5^{th} percentile (P5), 25^{th} percentile (P25), 75^{th} percentile (P75), and 95^{th} percentile (P95) of the variables of interest, as well as number of observations for each variable. Our main variable, *data scientist ratios*, is the total number of data scientists scaled by the total number of employees. The hiring rate is calculated as the ratio of newly hired data scientists to the average number of data scientists employed in the current and previous years. Hiring growth is measured as the logarithmic difference (i.e., growth rate) in the number of newly hired data scientists. The other variables include market capitalization (ME), book-to-market ratio (B/M), investment rate (I/K), return on assets (ROA), markup, profit margin, book leverage (Lev), operating leverage (OP Lev), and asset growth.[8]

We have a total of 29,462 firm-year observations with non-missing data scientist ratios. The average data ratio is 6.90%, suggesting that firms hire non-trivial amounts of employees with data skills in the labor force. Industry-level summary statistics for data scientist ratios are presented in Section C.2 of the Appendix.

Panel B of Table 1 presents a correlation matrix for all variables considered in Panel A. In the following subsection, we conduct the univariate portfolio sort, factor regressions, and Fama-MacBeth regressions to document the return predictability of data scientist ratios. In the rest of our analyses, we focus on data scientist ratios according to the general occupation descriptions.

3.3 Data Scientists and Firm Dynamics

This section examines the empirical relationship between data scientist ratios and firm characteristics. Specifically, we investigate the link between the data scientist ratio and market power, as well as the relationship between the data scientist ratio and firms' forecast accuracy.

[8] Detailed information on the markup construction refers to De Loecker et al. (2020).

Table 1: **Statistics and Correlations**

This table presents summary statistics in Panel A and a correlation matrix in Panel B for the firm-year sample. The data scientist ratio (DS Ratio) is measured as the percentage of data scientists among all employees hired by a firm, calculated as the number of data scientists divided by the total number of data scientists and non-data scientists. The hiring rate (DS Hiring Rate) is calculated as the ratio of newly hired data scientists to the average number of data scientists employed in the current and previous years. The hiring growth (DS Hiring Growth) is measured as the percentage growth in newly hired data scientists, calculated as the logarithmic difference in their numbers across consecutive years. ME is market capitalization deflated by CPI (measured in 2009 million USD) at the end of June. B/M is the ratio of book equity to market capitalization. I/K is capital expenditures (item CAPX) divided by property, plant, and equipment (item PPENT). Return on assets (ROA) is the gross profit (item GP) scaled by total assets (item AT). Markup is used to measure the market power, following De Loecker et al. (2020). Profit margin is sales revenue (SALE) minus the cost of goods sold (item COGS) and then divided by the cost of goods sold. Book leverage (Lev) is the summation of current liabilities (item DLC) and long-term debt (item DLTT) scaled by total assets. Operating leverage (OP Lev) is the selling, general, and administrative expenses (item XSGA) scaled by gross property, plant, and equipment (item PPEGT). Asset growth is measured as the logarithmic difference (i.e., growth rate) in total assets deflated by CPI. We report the pooled mean, standard deviation (Std), 5^{th} percentile (P5), 25^{th} percentile (P25), median, 75^{th} percentile (P75), and 95^{th} percentile (P95). Observations denote the valid number of observations for each variable. The sample period is 2008 to 2021 at an annual frequency.

	DS Ratio	**DS Hiring Rate**	**DS Hiring Growth**	**ME**	**B/M**	**I/K**	**GP/AT**	**Markup**	**Margin**	**Lev**	**OP Lev**	**Asset Growth**
					Panel A: Summary Statistics							
Mean	6.90	0.23	4.67	6,833.55	0.65	0.29	0.31	1.87	1.56	0.25	1.87	4.69
Std	7.13	0.22	245.17	34,608.93	0.81	4.58	0.33	1.35	13.31	0.26	10.53	29.00
P5	0.43	0.00	-260.54	18.25	0.08	0.04	-0.10	0.97	-0.29	0.00	0.00	-32.26
P25	2.48	0.10	-37.4	155.73	0.26	0.11	0.16	1.20	0.28	0.03	0.12	-5.96
Median	4.44	0.18	4.15	763.45	0.47	0.19	0.29	1.44	0.57	0.21	0.58	2.45
P75	8.73	0.29	41.61	3,144.98	0.80	0.32	0.45	1.96	1.26	0.37	1.59	12.38
P95	21.00	0.64	218.22	25,581.19	1.69	0.63	0.81	4.30	4.87	0.66	5.86	50.10
Observations	29,462	26,344	25,934	28,889	27,941	29,385	29,462	25,137	29,433	29,329	29,054	27,886
					Panel B: Correlation							
DS Ratio	1											
Hiring Rate	0.15	1										
Hiring Growth	0.01	0.38	1									
ME	0.12	0.01	0.00	1								
B/M	-0.16	-0.11	-0.04	-0.09	1							
I/K	0.01	0.01	0.00	0.00	-0.01	1						
GP/AT	0.04	0.03	0.01	-0.02	-0.19	0.00	1					
Markup	0.37	0.19	0.01	0.06	-0.2	0.01	0.32	1				
Margin	0.34	0.17	0.01	0.07	-0.16	0.01	0.25	0.95	1			
Lev	-0.15	-0.05	-0.01	0.04	0.01	-0.01	-0.26	-0.16	-0.11	1		
OP Lev	0.04	0.09	0.00	-0.02	-0.03	0.19	0.05	0.10	0.07	-0.07	1	
Asset Growth	0.07	0.14	0.03	0.03	-0.17	0.03	-0.03	0.12	0.11	0.05	0.05	1

3.3.1 Data Scientists and Markup

Previously, we have shown preliminary evidence on the positive correlation between data scientists hiring and market power in Panel B of Table 5. To further strengthen the linkage, we perform OLS regressions controlling for firm characteristics. Specifically, we run the following regression,

$$\text{Markup}_{i,t} = a_i + \delta_t + b \times \text{DS}_{i,t} + c \times Controls_{i,t} + \varepsilon_{i,t}, \tag{12}$$

where $\text{DS}_{i,t}$ is our proxy for data scientists, a_i represents industry or firm fixed effects, and δ_t denotes time fixed effects. We use three proxies for the hiring of data scientists: the data scientist ratio (also used in the univariate portfolio sort exercise in Table 5), the data scientist hiring rate, and the growth rate of newly hired data scientists. Control variables include firm size, book-to-market ratio (B/M), investment-to-capital ratio (I/K), profitability, book leverage (Lev), selling, general, and administrative (SGA) expenses to total assets ratio, and asset growth.

Specification 1 of Table 2 shows that a one-standard-deviation increase in the data scientist ratio or hiring rates is associated with a 15% increase in markup. This relationship holds with alternative measures, such as the data scientist hiring growth rate, where a one-standard-deviation increase corresponds to a 1.6% to 3.9% rise in markup. Specification 4 provides further robustness results, the positive correlation is also significant after controlling for firm-level fixed effects. We also control for variables strongly correlated with markup, as suggested by De Loecker et al. (2020), such as size and SGA. This empirical evidence strongly supports the notion that data scientists contribute to firm markups, even after accounting for other firm characteristics.

3.3.2 Data Scientists and Forecast Errors

We hypothesize that data technology enhances the information quality of a firm to understand consumer behaviors better and charge higher markups. This section tests the link between data technology and information quality. Information quality is not directly observed, so we employ sales forecast errors as proxies. Specifically, we investigate whether firms with more data scientists exhibit lower forecast errors, focusing primarily on management's sales forecast errors. The rationale is that data scientists can process data, providing managers with better insights into consumer demand, which in turn helps managers make more accurate sales forecasts. Management forecasts offer a clear advantage over analyst forecasts, as they incorporate information generated within the firm and publicly available data that analysts can access. We focus on sales rather than earnings per share because sales can capture consumers' demand better.

Table 2: **Markup and Data Scientists**

This table examines the relationship between firm markups and data scientist-related measures. The variables of interest are the DS ratio, DS hiring rate, and DS hiring growth, all of which are standardized to have a mean of zero and a standard deviation of one for comparability. We report panel regressions of markups on data scientist-related measures, along with other firm characteristics, as in Equation (12). Standard errors are double clustered at the firm and year level.

	(1)	**(2)**	**(3)**	**(4)**
DS Ratio	0.155***			
	(3.696)			
DS Hiring Rate		0.148***		
		(7.061)		
DS Hiring Growth			0.039**	0.016*
			(2.464)	(2.039)
Log ME	0.062***	0.070***	0.073***	0.251***
	(4.246)	(4.701)	(4.768)	(5.365)
B/M	-0.017*	-0.017	-0.023	-0.002
	(-1.794)	(-1.659)	(-1.770)	(-0.707)
I/K	-0.001	-0.003	-0.003	0.001
	(-1.615)	(-1.010)	(-1.055)	(0.420)
GP/AT	1.992***	1.977***	1.894***	2.545***
	(14.982)	(13.538)	(13.757)	(14.409)
Lev	-0.063	-0.167	-0.188	0.002
	(-0.542)	(-1.316)	(-1.478)	(0.013)
SGA/AT	1.083***	1.156***	1.230***	0.669***
	(6.451)	(7.266)	(6.975)	(3.077)
Asset Growth	0.127***	0.113**	0.122***	0.016
	(3.179)	(2.965)	(3.074)	(0.908)
Observations	40,294	37,043	36,454	35,807
R-squared	0.180	0.183	0.181	0.746
Industry FE	Yes	Yes	Yes	No
Firm FE	No	No	No	Yes
Year FE	Yes	Yes	Yes	Yes
Cluster Year	Yes	Yes	Yes	Yes
Cluster Firm	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

We collect managerial annual sales estimates reported in the IBES Guidance database over our sample period. To quantify the precision of managerial forecasts, we construct a measure of revision improvement (RI) based on forecast updates, calculated as the logarithmic difference in absolute forecast errors between the beginning and end of the year. This approach is based on the assumption that firms with better data technology and forecasting capabilities more effectively resolve sales uncertainty as new information becomes available throughout the calendar year, resulting in improved year-end predictions,

$$\mathrm{RI}_{i,t} = -\log\left(|\mathrm{Forecast}_{i,t}^{\mathrm{last}} - \mathrm{Actual}_{i,t}|\right) + \log\left(|\mathrm{Forecast}_{i,t}^{\mathrm{first}} - \mathrm{Actual}_{i,t}|\right), \tag{13}$$

where $\mathrm{Forecast}_{i,t}^{\mathrm{last}}$ represents the value of the final managerial sales forecast from the IBES Guidance database at the end of the year, $\mathrm{Actual}_{i,t}$ denotes the actual sales reported at the end of the reporting period, and $\mathrm{Forecast}_{i,t}^{\mathrm{first}}$ refers to the initial managerial sales forecast at the beginning of the year. This measure captures the improvement in forecasting precision within a certain horizon, reflecting the firm's ability to refine its predictions as more information becomes available.[9] It also captures the idea that firms with more data scientists can quickly learn and incorporate all recent information.

To confirm that our data scientist ratio is a valid proxy for firms' data investment to improve learning, we examine whether firms with more data scientists experience reductions in forecast revisions. We test the relation between firm-level revision improvement and data scientist ratios by estimating the following OLS regression,

$$\mathrm{RI}_{i,t+h} = a_i + \delta_t + b \times \text{DS Hiring Rate}_{i,t} + c \times Controls_{i,t} + \varepsilon_{i,t}, \tag{14}$$

where $\mathrm{DS}_{i,t}$ is the proxy for data scientists, a_i represents industry or firm fixed effects, and δ_t denotes time fixed effects. We control for a rich set of firm characteristics, including size, B/M, I/K, profitability, book leverage, SGA over total asset ratio. Moreover, managers can potentially better forecast their own firms' performance because there are more analysts cover their firms. To address this issue, we also include analyst coverage regressions. Standard errors are double clustered at firm and year level.

Specifications 1 and 2 of Table 3 show that the estimated coefficient on the data scientist ratio b is significantly positive, demonstrating that firms with more data scientists can achieve better sales forecast accuracy contemporaneously. Specifications 3 and 4 demonstrate that hiring more data scientists also improves future forecast revisions. Given the persistent nature of data scientist hiring decisions, this suggests that firms consistently increasing their data scientist workforce maintain a persistent advantage in sales forecasting.

[9] Feng et al. (2009) also use forecast revision to proxy for informativeness.

Table 3: **Forecast Errors Regressions**

This table shows the link between the hiring rate with the forecast revision improvement. We report the results of the panel regressions Equation (13). Both revision improvement (RI) and data scientist hiring rate (DS hiring rate) are in percentage numbers. Industry fixed effects are based on NAICS two-digit industry classifications. Standard errors are double clustered at firm and year level. All regressions are conducted at the annual frequency.

	(1)	**(2)**	**(3)**	**(4)**
	$\mathrm{RI}_{i,t}$	$\mathrm{RI}_{i,t}$	$\mathrm{RI}_{i,t+1}$	$\mathrm{RI}_{i,t+1}$
DS hiring rate	0.215*	0.318*	0.219*	0.352***
	(1.792)	(1.819)	(1.888)	(3.268)
Log ME	0.019	-0.103*	0.013	0.055
	(1.331)	(-2.015)	(0.700)	(1.238)
B/M	0.079*	0.056	0.142***	0.155*
	(1.778)	(0.951)	(3.189)	(1.902)
I/K	-0.463***	-0.246**	-0.234***	-0.024
	(-3.809)	(-2.892)	(-3.108)	(-0.247)
GP/AT	-0.353***	-0.132	-0.284***	0.233
	(-3.502)	(-0.604)	(-3.099)	(0.841)
Lev	-0.180	-0.177	-0.011	0.138
	(-1.272)	(-0.919)	(-0.071)	(0.704)
SGA/AT	0.219**	0.267	0.164*	0.110
	(2.438)	(0.973)	(2.119)	(0.388)
Number of Analysts	-0.005*	-0.001	-0.003	0.008
	(-2.012)	(-0.195)	(-0.956)	(1.677)
Observations	10,435	10,046	10,490	10,091
R-squared	0.032	0.233	0.027	0.226
Industry FE	Yes	No	Yes	No
Firm FE	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes
Cluster Year	Yes	Yes	Yes	Yes
Cluster Firm	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

3.3.3 Data Scientists and Job Requirements

To empirically investigate the role of data scientists in firms, we examine whether firms with higher data scientist ratios also exhibit greater demand for data-related technical skills and tasks. Specifically, we construct a job requirement measure using keywords extracted from job descriptions in the Lightcast dataset, categorizing them into four groups: Technical, Forecast, Pricing, and Data. The first two reflect skill-related terms, while the latter two refer to job responsibilities. The Technical category includes keywords related to computer science skills, such as machine learning, random forest, and SVM. The Forecast category covers forecasting-related terms, such as causal analysis and regression. The Pricing category includes keywords associated with pricing tasks, such as consumer demand, price elasticity, and spending pattern. The Data category contains keywords related to data analysis tasks, such as data science and data analytics.

For each job posting, we compute the frequency of keywords in each category, normalized by the total number of words in the job description. We then average this normalized measure at the firm-year level.

Table 4 reports the regression results. We regress the firm's data scientist ratio on these job requirement measures. The main independent variable, denoted *Requirements*, is the normalized frequency of job task keywords from firm i's postings in year t. We control for key firm fundamentals, including firm size, book-to-market ratio, investment rate (I/K), profitability, leverage, SG&A-to-asset ratio, and asset growth. Across all specifications, the job requirement measures are positively and significantly associated with the data scientist ratio. A one-standard-deviation increase in the job requirement measure is associated with an approximate 0.05 standard deviation increase in the data scientist ratio. This is translated into 0.4% increase in the data scientist ratio, in comparison to its median of 5%. We have also performed panel regression with firm and year fixed effects, the results are quantitatively similar. These findings suggest that firms hiring more data scientists place greater emphasis on technical skills, forecasting ability, and job tasks related to pricing and data analysis. This supports the interpretation of the data scientist ratio as a valid proxy for firm-level investments in data capabilities aimed at learning unobserved consumer demand.

3.4 Data Scientists and Stock Returns

3.4.1 Univariate Portfolio Sorting

To explore the connection between data scientist ratios and expected stock returns in the cross-section, we construct five portfolios by sorting firms based on their data scientist ratio. The data scientist ratio is calculated as the year-end number of data scientists divided by the total number of employees. These portfolios are rebalanced annually.

At the end of each June in year t, we sort firms into quintile portfolios based on their data scientist ratios at the end of year $t-1$. Additionally, the portfolio sort is relative to peers within their NAICS two-digit industries, using NYSE breaking points. Firms with non-missing data scientist ratios in year $t-1$ are allocated to portfolios, with the low (high) portfolio comprising firms with the lowest (highest) ratios. To examine the relationship

Table 4: **Data Scientist Ratios and Job Requirements**

This table examines the relationship between data scientist ratios and job requirements. We extract keywords from job descriptions in the Lightcast data and classify them into four categories according to job skill and task requirements: Technical, Forecast, Pricing, and Data. We estimate panel regressions by regressing firm's data scientist ratio on normalized job requirements, controlling for firm characteristics, industry, and year fixed effects. Job requirements are measured as the frequency of the corresponding keywords in each category, normalized by the total number of words in the job description, they are then averaged for by firm and year. All independent variables are normalized to zero mean and unit standard deviation after winsorization at the 1st and 99th percentiles to reduce the impact of outliers. The t-statistics are based on standard errors that are double clustered by firm and year. All regressions are conducted at an annual frequency. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

	(1)	(2)	(3)	(4)
	Technical	Forecast	Pricing	Data
Requirements	0.05**	0.05**	0.03*	0.05**
	(2.47)	(2.25)	(1.81)	(2.52)
Log(ME)	0.16***	0.15***	0.17***	0.15***
	(5.05)	(4.85)	(5.71)	(4.95)
B/M	-0.07**	-0.07***	-0.07***	-0.07***
	(-2.56)	(-2.59)	(-2.59)	(-2.63)
I/K	0.10***	0.11***	0.10***	0.10***
	(4.49)	(4.53)	(4.56)	(4.47)
GP/AT	-0.08***	-0.08***	-0.08***	-0.08***
	(-3.12)	(-3.05)	(-3.15)	(-3.07)
Lev	-0.07**	-0.07**	-0.07***	-0.07**
	(-2.41)	(-2.50)	(-2.59)	(-2.47)
SGA/AT	0.03	0.02	0.02	0.02
	(0.59)	(0.49)	(0.42)	(0.49)
Asset Growth	0.01	0.01	0.01	0.01
	(1.34)	(1.29)	(1.10)	(1.35)
Observations	3,765	3,765	3,765	3,765
R^2	0.54	0.54	0.54	0.54
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

between data scientist ratios and returns, we construct a high-minus-low portfolio by taking a long position in the highest quintile and a short position in the lowest.

We calculate value-weighted monthly returns for these portfolios from July of year t to June of year $t+1$, weighting firms by market capitalization at portfolio formation to reduce the impact of small, less tradable firms. Firms with asset values or sales below $1 million are excluded to further minimize the influence of very small firms.

In Panels A of Table 5, the top row presents the *annualized* average excess stock return in percentage (E[R]-$\mathrm{R_f}$, in excess of the risk-free rate), t-statistic, standard deviation, and Sharpe ratio for the six portfolios we consider. The table shows that a firm's data scientist ratio forecasts stock returns. Taking Panel A, which uses data scientist ratio, as an example, the quintile portfolio sorts from low to high have annualized excess returns of 11.21%, 12.75%, 12.32%, 13.28%, and 15.34%, respectively. More importantly, the H-L portfolio has an annualized excess return of 4.13% with a t-statistic of 2.00. In addition, the Sharpe ratios of the quintile portfolios are 0.67, 0.83, 0.85, 0.89, and 1.05, respectively, and that of the high-minus-low portfolio is 0.45, which is comparable to the Sharpe ratio of the aggregate equity premium. The finding that the return on the H-L portfolio is economically large and statistically significant suggests a significant predictive ability of firm-level data scientist ratios for stock returns. Overall, Table 5 provides empirical evidence that firm-level data scientist ratios help explain subsequent stock returns.

Panels B of Table 5 reports the average firm characteristics across quintile portfolios. On average, firms in the highest quintile group exhibit a data scientist ratio of 16.58% per year, compared to just 0.01% per year for firms in the lowest quintile group. Additionally, firms with higher data scientist ratios tend to have higher data scientist hiring rates and growth rates. Furthermore, we find that firms with high data scientist hiring ratios are characterized by lower book-to-market ratios but higher investment rates, markup, profit margins, operating leverage, as well as asset growth. However, there is little variation in book leverage and profitability across the quintile-sorted portfolios.

3.4.2 Asset Pricing Factor Regressions

Next, we analyze the extent to which the variability in the average returns of the data-scientist-hiring-rates-sorted portfolios can be explained by exposure to standard risk factors as proposed by the Fama and French (2015) five-factor model, the Hou, Xue, and Zhang (2015) q-factor model.[10]

[10] The Fama and French factors are sourced from Kenneth French's data library (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). The HXZ factors are obtained from the q-factors data library (http://globalq.org/index.html).

Table 5: **Univariate Portfolio Sorting**

This table reports average excess returns for five portfolios sorted on data scientist ratios, using NYSE breaking points, relative to their industry peers in Panel A and the time-series average of the cross-sectional medians of firm characteristics in Panel B, for which we use the NAICS two-digit industry classifications, and rebalance portfolios at the end of every June. The sample starts from July 2009 to December 2022, financial firms are excluded. We report average excess returns over the risk-free rate (E[R]-R_f), *t*-statistics, and Sharpe ratios (SR) across five portfolios in each panel. Portfolio returns are value-weighted by firms' market capitalization, and are multiplied by 12 to make the magnitude comparable to annualized returns. *t*-statistics based on standard errors using the Newey-West correction are reported in parentheses. The definitions of firm characteristics are described in Table 1.

Panel A: Univariate Sorted Portfolios

	L	**2**	**3**	**4**	**H**	**H-L**
E[R] - R_f (%)	11.21	12.75	12.32	13.28	15.34	4.13
[t]	3.24	3.91	3.88	4.35	5.27	2.00
SR	0.67	0.83	0.85	0.89	1.05	0.45

Panel B: Firm Characteristics

	L	**2**	**3**	**4**	**H**
DS Ratio (%)	0.01	2.13	4.04	7.20	16.58
Hiring Rate	0.14	0.17	0.17	0.19	0.20
Hiring Growth (%)	-1.15	2.86	3.93	4.86	5.46
Log ME	5.83	6.84	7.00	7.18	6.95
BM	0.57	0.53	0.48	0.41	0.39
I/K	0.13	0.14	0.15	0.16	0.16
GP/AT	0.26	0.29	0.31	0.32	0.29
Markup	1.31	1.36	1.42	1.52	1.47
Profit Margin	0.46	0.50	0.59	0.68	0.70
Book Lev	0.22	0.25	0.23	0.21	0.16
OP Lev	0.38	0.38	0.48	0.64	0.76
Asset Growth (%)	1.87	2.04	2.54	2.44	3.47

To test the standard risk factor models, we perform time-series regressions of data-scientist-hiring-rates-sorted portfolios' excess returns on the Fama and French (2015) five-factor model (the market factor-MKT, the size factor-SMB, the value factor-HML, the profitability factor-RMW, the investment factor-CMA) in Panel A and on the Hou, Xue, and Zhang (2015) q-factor model (the market factor-MKT, the size factor-SMB, the investment factor-I/A, the profitability factor-ROE) in Panel B, respectively. Such time-series regressions enable us to estimate the betas (i.e., risk exposures) of each portfolio's excess return on various risk factors and to estimate each portfolio's risk-adjusted return (i.e., alphas in

%). We annualize the excess returns and alphas in Table 6.

Table 6: **Asset Pricing Factor Tests**

This table shows asset pricing factor tests for five portfolios sorted on data scientist ratios, using NYSE break points, relative to their industry peers, for which we use the NAICS two-digit industry classifications and rebalance portfolios at the end of every June. The results reflect monthly data, for which the sample starts from July 2009 to December 2022. To adjust for risk exposure, we perform time-series regressions of data-scientist-hiring-rates-sorted portfolios' excess returns on the Fama and French (2015) five factors (MKT, the size factor-SMB, the value factor-HML, the profitability factor-RMW, and the investment factor-CMA) in Panel A and on the Hou, Xue, and Zhang (2015) q-factors (MKT, SMB, the investment factor-I/A, and the profitability factor-ROE) in Panel B, respectively. Data on the Fama-French five-factor is from Kenneth French's website. Data on the I/A and ROE factors are provided by the q data library. These betas, together with alphas, are annualized by multiplying 12. The Newey-West adjusted *t* statistics are reported in parentheses.

	L	**2**	**3**	**4**	**H**	**H-L**
			Panel A: FF5			
α_{FF5}	-3.05	-1.73	-0.53	-0.91	1.76	4.81
[t]	-2.89	-1.14	-0.48	-1.25	2.21	3.03
MKT	0.99	0.98	0.91	1.00	0.99	-0.00
[t]	46.88	36.19	47.36	49.27	58.14	-0.09
SMB	0.35	0.18	0.12	0.00	-0.13	-0.48
[t]	8.26	4.56	3.25	0.10	-3.90	-7.45
HML	0.08	-0.01	0.04	-0.08	-0.20	-0.28
[t]	1.70	-0.13	0.59	-2.09	-6.39	-3.46
RMW	0.14	0.21	0.20	0.10	0.08	-0.05
[t]	1.88	2.99	2.82	2.52	2.38	-0.66
CMA	-0.02	0.23	0.01	0.15	-0.02	-0.01
[t]	-0.11	1.45	0.08	2.47	-0.25	-0.03
			Panel B: HXZ			
α_{HXZ}	-1.45	-0.28	0.72	-0.34	1.68	3.13
[t]	-1.36	-0.18	0.76	-0.41	4.40	3.32
MKT	0.99	0.99	0.91	0.99	1.00	0.01
[t]	56.72	44.56	32.05	53.37	89.45	0.59
SMB	0.32	0.11	0.10	-0.07	-0.22	-0.54
[t]	6.78	1.78	2.41	-1.66	-8.06	-12.68
I/A	0.09	0.24	0.08	0.06	-0.22	-0.31
[t]	0.94	2.49	1.20	1.15	-4.50	-2.81
ROE	-0.02	0.03	0.03	-0.04	0.06	0.08
[t]	-0.38	0.47	0.85	-0.68	1.73	1.04

As presented in Table 6, the risk-adjusted returns (intercepts) of the high-minus-low portfolio sorted by data scientist ratios remain notably large and statistically significant. These intercepts range from 4.81% for the Fama and French (2015) five-factor model in Panel A to 3.13% for the Hou, Xue, and Zhang (2015) q-factor model in Panel B. These intercepts are all at least 3.03 standard errors above zero, indicating high statistical significance. Additionally, the alphas estimated by both the Fama-French five-factor model and the HXZ q-factor model remain comparable to the return spread observed in the univariate sorting (Table 5). Furthermore, the high-minus-low portfolio's returns exhibit significantly negative size and value betas in relation to both the Fama and French (2015) five-factor model and the Hou, Xue, and Zhang (2015) q-factor model. However, these returns show insignificant market betas with respect to both models.

Taken altogether, the outcomes from the asset pricing factor tests detailed in Table 6 indicate that the variation in cross-sectional returns among portfolios categorized by data scientist ratios cannot be absorbed by the Fama French five-factor model (Fama and French (2015)) and the HXZ q-factor model (Hou, Xue, and Zhang (2015)). Consequently, the elevated returns linked to data scientist ratios are not explained by common risk factors. In the subsequent subsection, we reinforce the association between data scientist ratios and returns by utilizing Fama-Macbeth regressions.

3.4.3 Fama-Macbeth Regressions

We further investigate the predictability of data scientist ratios for cross-sectional stock returns using Fama-MacBeth cross-sectional regressions (Fama and MacBeth (1973)). This regression enables us to account for an extensive array of firm characteristics that predict stock returns. Moreover, it allows us to explore whether the positive relationship between data scientist ratios and returns can be attributed to other established predictors at the firm level in the literature.

We perform cross-sectional regressions for each month spanning from July of year t to June of year $t+1$ as expressed in the following equation:

$$R_{i,t+1} - R_{f,t+1} \quad = \quad a_j + b \times \text{DS Ratio}_{i,t} + c \times Control_{i,t} + \varepsilon_{it}, \tag{15}$$

where a_j captures the industry fixed effects. Within each month, we regress the monthly returns of individual stocks (annualized by multiplying by 12) on the data scientist ratios of year $t-1$, with control variables known by the end of June of year t, and industry fixed effects at two-digit NAICS level. The control variables include the logarithm of market capitalization at the end of each June (Size), which is deflated by the CPI index, book-to-market ratio (B/M), investment rate (I/K), profitability (GP/AT), book leverage (Lev), operating leverage (OP Lev), SGA ratio (SGA/AT), markup, and asset growth. To mitigate the impact of outliers, all independent variables are winsorized at the 1st and 99th percentiles. Additionally, they are normalized to have a mean of zero and a standard deviation of one to facilitate comparisons.

Table 7: **Fama-Macbeth Regressions**

This table presents the results of Fama-MacBeth regressions, in which we analyze individual stock excess returns based on their data scientist ratios and alternative variables that are relevant in the literature. The regressions are conducted in a cross-sectional manner for each month, spanning from July of year t to June of year $t+1$. Specifically, in each month, we regress the monthly excess returns of individual stocks (annualized by multiplying by 12) on the data scientist ratio from year $t-1$, various sets of control variables known by the end of June of year t, and industry fixed effects. Industry categories are defined using NAIC two-digit industry classifications. All independent variables are normalized to have a zero mean and a one-standard-deviation, after winsorization at the 1st and 99th percentiles. The reported t-statistics are computed based on standard errors estimated using the Newey-West correction. The sample period for the analysis spans from July 2009 to December 2022.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
DS Ratio	2.07**	2.21***	2.02**	2.31***	1.88**	2.28***
	(2.48)	(2.77)	(2.40)	(2.91)	(2.23)	(2.74)
Log ME	-0.26	-0.31	-0.82	-1.78	-0.10	-0.39
	(-0.16)	(-0.19)	(-0.54)	(-1.40)	(-0.06)	(-0.25)
B/M	0.91	0.84	0.67	0.24	0.77	0.39
	(0.95)	(0.88)	(0.70)	(0.26)	(0.73)	(0.41)
I/K	-1.65**	-1.49**	-0.92	-1.49**	-2.37***	-1.50**
	(-2.25)	(-2.09)	(-1.31)	(-2.12)	(-3.18)	(-2.07)
GP/AT	1.88*	1.96*	2.45**	4.95**	3.33***	1.85*
	(1.75)	(1.84)	(2.15)	(2.53)	(2.63)	(1.77)
Lev		0.85				
		(0.87)				
OP Lev			-2.75***			
			(-3.35)			
SGA/AT				-4.56***		
				(-2.72)		
Markup					-0.63	
					(-0.96)	
Asset Growth						-1.98***
						(-2.93)
Observations	335,255	333,672	331,177	335,255	287,116	287,116
R-squared	0.06	0.06	0.06	0.06	0.06	0.06
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes

In Table 7, we report the average slopes from monthly regressions, and the corresponding t-statistics are the average slopes divided by their time-series standard errors. We annualize the slopes. The results support the return predictive ability of data scientist ratios. In Specification 1, data scientist ratios significantly positively predict future stock returns with a slope coefficient of 2.07 and a t-statistic of 2.48. This finding is consistent with the predictability and implies that a one-standard-deviation increase in data scientist ratios leads to a significant increase in the annualized stock return of 2.07%.

The results of these regressions are consistent with the results of the univariate portfolio sorting on data scientist ratios in Table 5, which show that data scientist ratios significantly positively predict future stock returns. From Specifications 2 to 6, data scientist ratios positively predict stock returns with statistically significant slope coefficients when we include control variables known to predict stock returns in the cross-section: size, book-to-market ratio, investment rate, profitability, book leverage, operating leverage, markup, and asset growth. More importantly, the predictability of data scientist ratios is not subsumed by known predictors of stock returns in the literature, even when we include all control variables jointly to run a horse race from Specifications 2 to 6.

4 Quantitative Results

In previous section, we document a positive association between information quality, data scientists and markups. Furthermore, we find evidence that the data scientist induced operating leverage effect dominates and firms with higher data scientist ratio earns higher stock returns. In this section, we explore the quantitative implications of our model. We calibrate our model and evaluate its ability to jointly account for the cross-section of firm characteristics and stock returns. The quantitative model yield additional insights on the cyclicality of data scientist hiring which we bring to the empirical analysis.

4.1 Calibration

The model is solved and calibrated at an annual frequency. We calibrate our model to match the firm dynamics whenever possible. The parameters are presented in Table 8.

The first block of parameters in Table 8 are related to the stochastic discount factor. We set β to match the level of risk-free rate at 1%. We set γ_0 and γ_1 to match the mean and volatility of the equity premium.

The second block of parameters is production technology parameters. Capital share α is set to 0.33 as in the business cycle literature. To match the average investment rates of

Table 8: **Calibration**

Parameter	Symbol	Values
Discount rate	β	0.9
Level of price of risk	γ_0	22.5
Sensitivity parameter of price of risk	γ_1	-15
Capital share in production	α	0.33
Return to scale	ϕ	0.85
Depreciation rate of capital	δ_K	0.15
Investment adjustment cost	c_K	1.5
Quit rate of data scientists	δ_N	0.15
Fixed cost	F	0.805
Data scientists adjustment cost	c_N	10
Level parmeter of data production	ν_0	1
Sensitivity parameter of data production	ν_1	0.65
Level parameter of data scientist's wage	τ_1^N	0.2
Sensitivity of the data scientist's wage to productivity	τ_2^N	0.65
Demand elasticity	η	8
Demand curve sensitivity to taste in steady state	$\bar{\Theta}$	0.63
Volatility of signal	σ_s	0.1
Volatility of forecasting technology	σ_θ	0.32
Persistence of TFP	ρ_A	0.91
Std of TFP shock	σ_A	0.02
Persistence of dispersion	ρ_Z	0.7
Std of dispersion shock	σ_Z	0.15

physical capital, we set the depreciation rate at 15%. The capital adjustment cost parameter c_K is set to match the cross-sectional volatility of investment rates. We set the attrition rate of data scientists at 15% as in the data. The parameter ν_0 is a scaling parameter that controls the data production; we normalize it to 1. The sensitivity parameter of data scientists in the data production function ν_1 is set to 0.65, roughly the same as the labor share in final goods production. The fixed cost parameter is set at 0.805 to match the average book-to-market ratio in the data, around 0.5.

For convex adjustment costs of physical capital investment G^K and data scientist G^N, we assume quadratic forms. For capital investment, we have

$$G^K(I_{it}, K_{it}) = \frac{c_K}{2}\left(\frac{I_{it}}{K_{it}} - \delta_K\right)^2$$

For data scientist stock, we have a similar quadratic form,

$$G^N(I_{it}, K_{it}) = \frac{c_N}{2}\left(\frac{H_{it}}{N_{it}} - \tilde{\delta}_N\right)^2$$

where c_K and c_N controls the severity of capital investment and hiring frictions. The anchoring parameter $\tilde{\delta}_N$ is chosen so that data scientist stock is positive in the steady state and stochastic simulations.

The wage process is specified as

$$W_t^N = \tau_1^N e^{\tau_2^N(a_t - \bar{a})}.$$

The scale parameter τ_1^N is calibrated to match the data scientists' labor share relative to sales. The aggregate labor share is around 66%, and data scientists consist of around 4% of the employees in our sample. Hence, their wage bill to total sales ratio should be around 2%. The sensitivity parameter τ_2^N is calibrated to be less sensitive than the skilled labor in Belo et al. (2017).

The third group of parameters relates to the demand curve. Following the business cycle literature, we set the demand elasticity, η, to 8. The sensitivity parameter to aggregate taste $\bar{\Theta}$ is set to 0.63 so that the average markup of the simulated economy aligns with the average markup observed in our sample. With ν_0 normalized to 1, the relative importance of data production technology is controlled by the precision of taste σ_θ^{-2} and the precision of the signal σ_s^{-2}. If the signal is relatively noisy, then firms need more data scientist to improve the quality of the learning by hiring more data scientists, which eventually increase the average employment of data scientists. Therefore, σ_θ can be interpreted as a scaling parameter, we set it together with $\bar{\Theta}$ to pin down the average markup of the economy. The volatility of signal σ_s is set to match the data scientists' wage bill to output ratio. The wage information of data scientists are not of high quality, but we observe the data scientists to total employment ratio, which is around 4%. Therefore, given the aggregate labor share is roughly 2/3, the data scientists' compensation to output ratio should be around 2% or higher. We set σ_s to roughly match the data scientists share in output at 1%.[11]

Finally, the last group of parameters is about productivity shocks. We specify the pro-

[11] We assume data scientists are not factor inputs. Therefore, their compensation relative to final sale ratio is very low in general.

ductivity processes in logs

$$a_{t+1} - \bar{a} = \rho_A(a_t - \bar{a}) + \sigma_A \varepsilon_t^A,$$
$$z_{t+1} - \bar{z} = \rho_Z(z_t - \bar{z}) + \sigma_Z \varepsilon_{it}^Z,$$

where ε_t^A and ε_{it}^Z are i.i.d. shocks. They are set following the production asset pricing literature, such as Zhang (2005).

We directly specify the stochastic discount factor without explicitly modeling the household's problem since we focus on the firm dynamics and stock returns in the cross section. The pricing kernel is given by

$$\log(M_{t,t+1}) = \log(\beta) - \gamma_t \sigma_A \varepsilon_t^A - \frac{1}{2}\gamma_t^2 \sigma_A^2, \tag{16}$$

where $M_{t,t+1}$ is the stochastic discount factor from time t to $t+1$, and ε_t^A is the innovation to the aggregate productivity shock a_t. The time-varying parameter γ_t controls the price of risk, it is given by

$$\gamma_t = \gamma_0 + \gamma_1(a_t - \bar{a}).$$

Given the parameters calibrated in Table 8, we simulate the model economy. The aggregate moments are reported in Table 9. The aggregate moments roughly match the empirical moments in the data.

Table 9: **Aggregate Moments**

	Data	Model
Volatility of aggregate output growth	0.02	0.02
Investment rate volatility	0.21	0.19
Average market excess return (%)	7.31	7.75
Volatility of market excess return (%)	21.56	17.65
Average real risk-free rate (%)	1.28	1.82
Volatility of real risk-free rate (%)	0.96	1.25

4.2 Model Mechanism

To illustrate the model mechanism, we plot the impulse response functions (IRFs) in response to positive aggregate and firm-level shocks, respectively.

Impulse responses to aggregate productivity shocks Figure 1 shows the impulse response functions upon a positive aggregate productivity shock. The model generates similar impulse responses on the supply side as in the standard q-theory model. A positive supply productivity shock is good news for production: the firm's output goes up, and it invests more to take advantage of higher productivity in the near term. The persistent increase in productivity raises future cash flows, which boosts the firm's value.

Conversely, we observe an opposite response in the firm's hiring decisions, where the firm reduces its labor force skilled in data, as illustrated in the right panel of Figure 1. Proposition 1 confirms that the firm's demand for data scientists reflects the behavior of markup dynamics. Specifically, when markup is countercyclical, the firm's demand for data scientists should similarly be countercyclical. This countercyclical nature of markup dynamics arises from the inelastic habit component of the demand curve. Under the demand curve (9), actual demand contains both an elastic component and an inelastic component influenced by historical consumer demands. Following a positive productivity shock, the elastic component of the demand curve (9) becomes more significant, effectively increasing the elasticity of demand. Because markup is inversely related to effective elasticity, a positive productivity shock reduces the markup by enhancing elasticity.

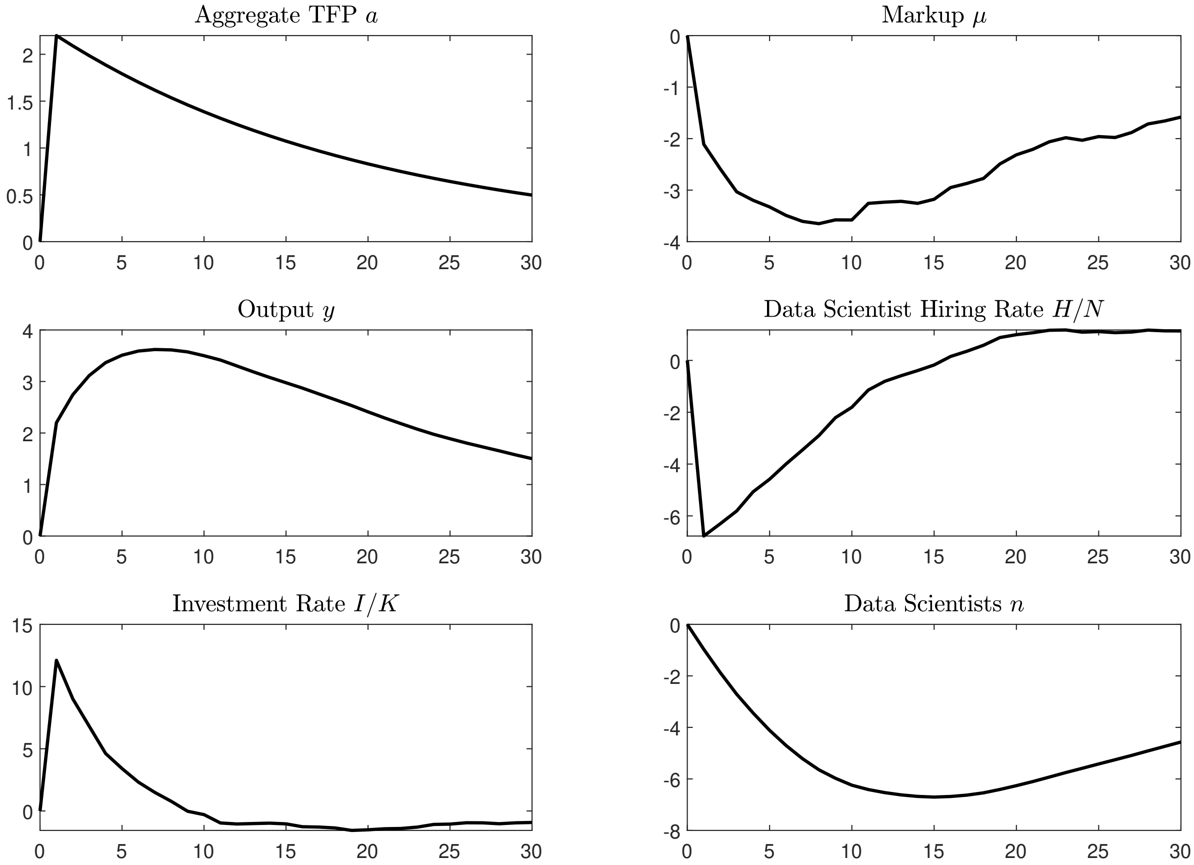
Impulse responses to firm-level productivity shocks The IRFs with respect to firm-level productivity shocks are shown in Figure 2. The left panels illustrate that the effect of a positive firm-level productivity shock aligns with the intuition of the standard q-theory model: output, investment, and firm value all increase when a firm experiences a positive firm-level productivity shock. In the right panels, the responses of markup and data scientist hiring rates are similar to those seen with an aggregate productivity shock, both decreasing following a positive firm-level productivity shock due to the countercyclical markup discussed in Proposition 3.

A firm-level productivity shock leads to an increase in precision. This difference is driven by two main channels. First, the data production technology, as seen in Equation (3), implies that past sales data as input for data production. Thus, a positive productivity shock helps the firm accumulate more data, leading to a delayed increase in precision Ω.

The second and more significant channel is shock persistence. As shown in Table 8, firm-level shocks are less persistent than aggregate shocks. As noted in Ai, Li, and Tong (2022), a less persistent shock encourages firms to focus on short-term cash flow rather than long-term investment. In our model, firms can boost short-term cash flow by increasing markup. Thus, in response to a transitory firm-level shock, firms reduce data scientist hiring and increase investment, but to a lesser extent than they would for a persistent aggregate

Figure 1: **Impulse Response to Aggregate Productivity Shocks**

This figure plots the log deviations from the steady-state for quantities and prices with respect to a one standard-deviation shock to the aggregate productivity shock. One period is a year. All parameters are calibrated as in Table 8.



shock. As illustrated in Figure 2, a 15% increase in firm-level productivity results in roughly a 20% decrease in data scientist hiring and a 20% increase in investment. In comparison, a 2.2% increase in aggregate TFP leads to a 10% drop in data scientist hiring and nearly a 15% rise in investment.

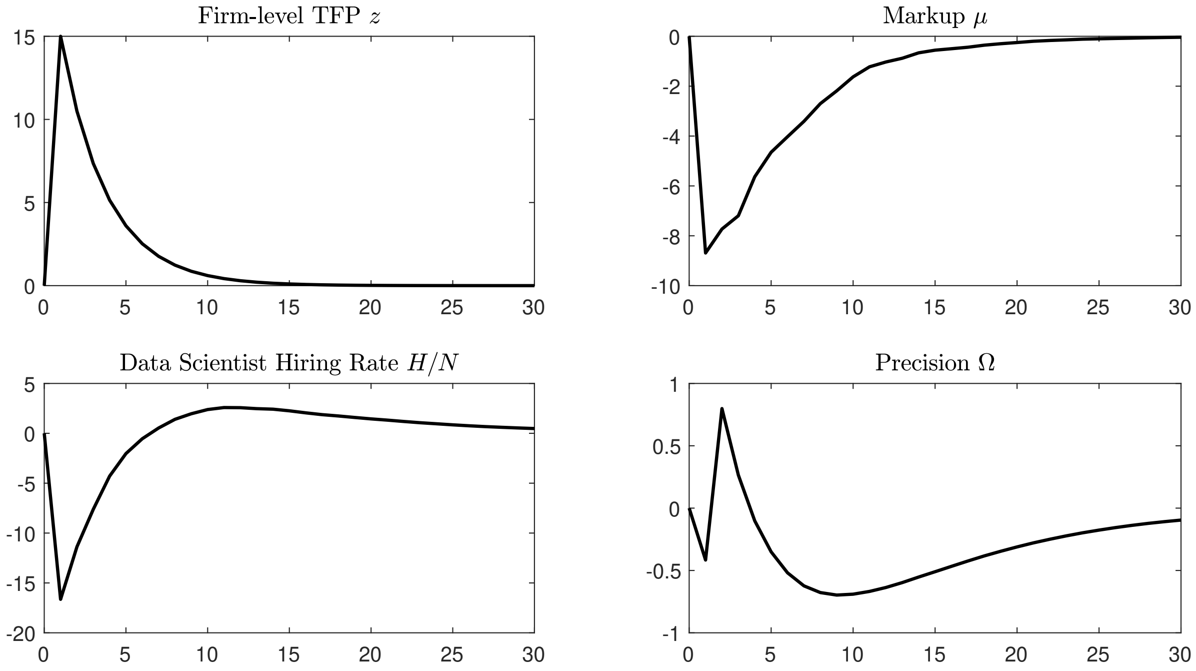
4.3 Implications for the Cross-Section

This section presents the model's implications on the cross-section's firm characteristics and stock returns. We simulate 4000 firms for 1000 periods. The number of firms roughly matches the sample size in CRSP.

We perform the same exercises as in the empirical analysis using the simulated data from the model. We sort firms into quintile portfolios each year based on their data scientists to total employee ratio, N. We report the statistics for data scientist ratio sorted portfolios.

Figure 2: **Impulse Response to Firm-Level Productivity Shocks**

This figure plots the log deviations from the steady-state for quantities and prices with respect to a one standard-deviation shock to the aggregate productivity shock. One period is a year. All parameters are calibrated as in Table 8.



The model replicates the empirical patterns documented in Section 3. Firms with more data scientists charge higher price markups, make more investments and earn higher returns.

The dynamics of markup and data scientists follow the economic intuition discussed in Section 4.2 and Proposition 3: the habit component in the demand curve leads to countercyclical markup and data scientists ratio, hence firms with more data scientists are relatively less productive firms. Firms with lower firm-level productivity would like to employ more data scientists because they improve the information quality of demand estimation, thus allowing firms to charge high markups.

Additionally, firms with more data scientists tend to invest slightly more on average. This is because, after the initial impact of a productivity shock, both the investment rate I/K and data scientist ratio N gradually return to their steady-state levels from below. This process creates a weak positive correlation between these two variables.

The model generates a substantial return spread across the data scientist ratio-sorted portfolios, as in Section 3. There are two economic forces drive this result. One is the hedging effect, as discussed in Section 4.2. Firms with low firm-level productivity employ more data scientists to boost sales. Therefore, more data scientists can provide a hedge against adverse productivity shocks. The other channel is the operating leverage channel.

Table 10: **Firm characteristics and expected returns: data and model**

This table compares the cross-sectional moments in the empirical data and the model simulated data at an annual frequency.

Panel A: Data

	L	**2**	**3**	**4**	**H**	**H-L**
DS Ratio (%)	0.00	2.13	4.04	7.20	16.58	
DS Hiring	0.14	0.17	0.17	0.19	0.20	
I/K	0.13	0.14	0.15	0.16	0.16	
Markup	1.31	1.36	1.42	1.52	1.47	
OP Lev	0.38	0.38	0.48	0.64	0.76	
Sale/Capital	0.49	0.59	0.69	0.82	0.86	
E[R] - R_f (%)	11.21	12.75	12.32	13.28	15.34	4.13

Panel B: Model

	L	**2**	**3**	**4**	**H**	**H-L**
DS Ratio (%)	3.81	4.22	4.59	5.06	6.19	
DS Hiring	0.14	0.14	0.14	0.14	0.14	
I/K	0.15	0.15	0.15	0.16	0.18	
Markup	1.25	1.32	1.42	1.58	1.96	
OP Lev	0.65	0.69	0.72	0.75	0.82	
Sale/Capital	0.86	0.89	0.94	1.01	1.19	
E[R] - R_f (%)	7.95	9.14	9.99	11.03	12.49	4.54

On the contrary, countercyclical data scientist hiring implies that firms are burdened with high labor compensation to data scientists in economic downturns. This data scientist-induced operating leverage oeffect is more prominent in firms experiencing low productivity, thus amplifying their exposures to aggregate risk. Additionally, firms face fixed cost, which makes less productive firms bear relatively higher operating leverage. Quantitatively, we find that the operating leverage channel dominates. As a result, firms with low firm-level productivity hire more data scientists and earn lower expected returns. Empirically, we find firms with more data scientists have higher operating leverage, as shown in Table 10.

4.4 Empirical Support for the Model Mechanism

In this section, we explore additional predictions of our model in the data by examining several key testable implications. As discussed in Section 4.3, our model predicts that firms increase data scientist hiring in economic downturns. This countercyclical hiring of data scientists raises firms' fixed costs, amplifiying the sensitivity of firms' cash flows to economic cycles to reflect a key mechanism of the operating leverage channel. To do so, we provide empirical evidence on the data scientist hiring at both aggregate and firm levels to validate the operating leverage channel.

Aggregate-level At the aggregate level, the short sample for data scientists makes identifying cyclical fluctuations long the time-series dimension challenging. We therefore use the ratio of aggregate investment related data processing to physical investment as a proxy for the aggregate data scientist ratio, based on the idea that data scientists require these equipments to perform their work. We estimate a VAR system with the following variables, ordered sequentially: log of aggregate TFP, log of total investment, log of IT investment share, log of consumption, log of GDP, and log of the GDP deflator.[12] We adopt the standard recursive ordering, such that TFP affects quantities immediately.

Figure 3 presents the IRFs of selected variables. A one-standard-deviation increase in aggregate TFP immediately raises aggregate physical investment. However, the ratio of IT investment to physical investment declines, reflecting the acyclicality of IT-related investments. This result suggests that IT investment relative to physical investment is countercyclical. This empirical observation aligns with the IRFs generated by our model, as in Figure 1.

Figure 3 suggests that IT investment relative to physical investment is countercyclical, as also evidenced by the significantly negative correlations between the HP-filtered ratio and its logarithmic difference with the GDP and TFP growth rates, respectively. During recessions, when GDP or TFP experiences a decline, IT-related investment tends to increase or decrease less than physical capital investment. This contributes to the observed rise in the ratio. Although it is a well-established fact that investment, in general, is procyclical, the distinctive cyclicality of the IT-to-general investment ratio points to potentially different economic mechanisms driving their dynamics across business cycles. As a robustness check, we estimated a VAR system using IT investment instead of IT investment share, as shown in

[12] The data for aggregate IT-related investment is the sum of lines 5, 78, 87, 88, and 93 in NIPA Table 2.7. Aggregate TFP growth rates are from the Federal Reserve Bank of San Francisco, and cumulative TFP growth rates are used to obtain the level of TFP. We have also tested a VAR model using growth rates of aggregate variables, yielding similar results.

Figure 3: **Time-series Patterns of the IT Investment Ratio**

This figure plots the impulse responses of selected variables to an orthogonoalized one-standard deviation shock to the aggregate TFP shock. The shaded bands represent the 95% confidence intervals calculated from 1000 bootstrap replications. The sample period is 1955 to 2022, at an annual frequency.

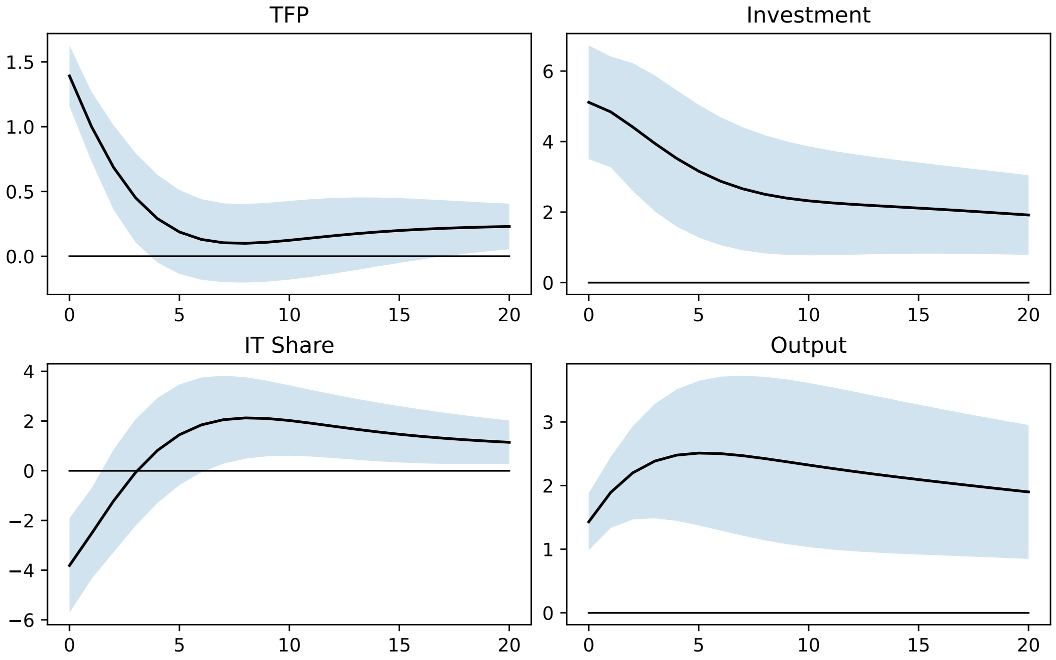


Figure B.1. The results confirm that IT investments remain acyclical, while physical investment is strongly procyclical. Consequently, the relative share of IT to physical investment responds negatively to aggregate productivity shocks.

Firm-level As shown Figure 2, firms cut data scientist hiring when they receive a positive firm-level productivity shock. To test this prediction, we regress variables related to data scientist hiring decisions on firm-level productivity shocks. The log productivity of firm-level TFP is from İmrohoroğlu and Tüzel (2014), we take the first difference as the TFP shock.

$$\text{DS}_{it} = a_i + \delta_t + b \times \Delta TFP_{it} + c \times Controls_{i,t} + \varepsilon_{i,t} \tag{17}$$

Consistent with our model predictions in Figure 2, we observe a significantly negative coefficient for both the data scientist ratio and the data scientist hiring rate. The results imply that when firms experience a positive productivity shock, they are would like to reduce their hiring of data scientists. These empirical findings strongly support our model prediction, which suggests that data hiring decisions are countercyclical.

Table 11: **Response to Firm-level Productivity Shock**

This table documents the relationship between firm-level productivity shocks and data scientist-related measures, specifically the data scientist ratio (DS ratio) and data scientist hiring rate (DS hiring rate). Both DS ratio and DS hiring rate are expressed as percentages. The growth rate of TFP (ΔTFP) is standardized to have a mean of zero and a standard deviation of one. Control variables include book leverage (Lev), book-to-market ratio (BM), investment rate (I/K), size (log of total assets), SGA-to-asset ratio (SGA/AT), and gross profitability. The regressions include both firm and year fixed effects. Standard errors are double clustered by the firm and year.

	(1)	(2)
	DS Ratio (%)	DS Hiring Rate (%)
ΔTFP	-0.016*	-0.147**
	(-1.858)	(-2.460)
Lev	0.164	-3.772**
	(1.301)	(-2.740)
BM	-0.003	-0.471**
	(-0.440)	(-2.674)
I/K	0.001	-0.007
	(1.239)	(-0.620)
Size	0.213***	3.620***
	(3.923)	(10.221)
SGA/AT	0.555*	-0.166
	(2.150)	(-0.050)
GPA	0.201	7.157**
	(0.876)	(2.871)
Observations	20,283	18,474
R-squared	0.981	0.500
Firm FE	Yes	Yes
Year FE	Yes	Yes
Cluster Year	Yes	Yes
Cluster Firm	Yes	Yes
Controls	Yes	Yes

5 Conclusion

In this paper, we investigate the role of data in firm dynamics and equity returns. We empirically demonstrate that firms with more data scientists have higher markups, lower sales forecast errors, and earn greater returns. To explain these empirical findings, we develop a heterogeneous firm model with endogenous data scientist hiring and learning decisions. In our model, firms face a downward-sloping demand curve due to their market power, which includes an unobservable inelastic consumer habit component. Firms hire data scientists to forecast this unobservable component. By improving their forecasts of consumer taste, firms can effectively increase the inelastic component of demand, allowing them to charge higher markups. Furthermore, our model suggests that the endogenous hiring of data scientists introduces an operating leverage channel, making firms with more data scientists riskier and, therefore, demanding higher returns. We also present additional empirical evidence that supports the mechanisms outlined in our model.

Future work could examine the risks associated with data investment through other well-known sources of aggregate fluctuations, such as the markup shock in Corhay, Li, and Tong (2022), the investment-specific technology shock as in Kogan and Papanikolaou (2013), or the financing shock as in Belo, Lin, and Yang (2019). The impact of data investment in general equilibrium is a promising future research area.

References

Abis, Simona, and Laura Veldkamp, 2024, The changing economics of knowledge production, *The Review of Financial Studies* 37, 89–118.

Agrawal, Ajay, John McHale, and Alex Oettl, 2018, Finding needles in haystacks: Artificial intelligence and recombinant growth, Working Paper 24541, National Bureau of Economic Research.

Aguerrevere, Felipe L, 2009, Real options, product market competition, and asset returns, *The Journal of Finance* 64, 957–983.

Ai, Hengjie, Jun E. Li, and Jincheng Tong, 2022, Equilibrium value and profitability premiums, *Working paper* .

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, 2024, Artificial intelligence, firm growth, and product innovation, *Journal of Financial Economics* 151, 103745.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp, 2018, Big data in finance and the growth of large firms, *Journal of Monetary Economics* 97, 71–87.

Belo, Frederico, Jun Li, Xiaoji Lin, and Xiaofei Zhao, 2017, Labor-force heterogeneity and asset prices: The importance of skilled labor, *Review of Financial Studies* 30, 3669–3709.

Belo, Frederico, Xiaoji Lin, and Fan Yang, 2019, External equity financing shocks, financial flows, and asset prices, forthcoming *Review of Financial Studies.*

Bian, Bo, Qiushi Huang, Ye Li, and Huan Tang, 2024, Data as a networked asset, *SSRN working paper* .

Brynjolfsson, Erik, and Kristina McElheran, 2019, Data in action: data-driven decision making and predictive analytics in us manufacturing, *Rotman School of Management Working Paper* .

Bustamante, M Cecilia, and Andres Donangelo, 2017, Product market competition and industry returns, *The Review of Financial Studies* 30, 4216–4266.

Corhay, Alexandre, Howard Kung, and Lukas Schmid, 2020, Competition, Markups, and Predictable Returns, *Review of Financial Studies* 33.

Corhay, Alexandre, Jun E Li, and Jincheng Tong, 2022, Markup shocks and asset prices, *Available at SSRN 4060403* .

Crouzet, Nicolas, and Janice Eberly, 2023, Rents and intangible capital: A q+ framework, *The Journal of Finance* 78, 1873–1916.

Crouzet, Nicolas, and Neil R. Mehrotra, 2020, Small and large firms over the business cycle, *American Economic Review* 110, 3549–3601.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, 2020, The rise of market power and the macroeconomic implications, *The Quarterly Journal of Economics* 135, 561–644.

Donangelo, Andres, François Gourio, Matthias Kehrig, and Miguel Palacios, 2019, The cross-section of labor leverage and equity returns, *Journal of Financial Economics* 132, 497–518.

Dou, Winston Wei, and Yan Ji, 2021, External financing and customer capital: A financial theory of markups, *Management Science* 67, 5569–5585.

Dou, Winston Wei, Yan Ji, and Wei Wu, 2021, Competition, profitability, and discount rates, *Journal of Financial Economics* 140, 582–620.

Eeckhout, Jan, and Laura Veldkamp, 2022, Data and market power, Technical report, National Bureau of Economic Research.

Eisfeldt, Andrea L., and Dimitris Papanikolaou, 2013, Organization capital and the cross-section of expected returns, *Journal of Finance* 68, 1365–1406.

Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of financial economics* 116, 1–22.

Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of political economy* 81, 607–636.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp, 2019, Big data and firm dynamics, in *AEA papers and proceedings*, volume 109, 38–42.

Farboodi, Maryam, and Laura Veldkamp, 2021, A model of the data economy, *NBER working paper* .

Feng, Mei, Chan Li, and Sarah McVay, 2009, Internal control and management guidance, *Journal of accounting and economics* 48, 190–209.

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek, 2017, Inflation dynamics during the financial crisis, *American Economic Review* 107, 785–823.

Heyerdahl-Larsen, Christian, 2014, Asset prices and real exchange rates with deep habits, *The Review of Financial Studies* 27, 3280–3317.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *The Review of Financial Studies* 28, 650–705.

İmrohoroğlu, Ayşe, and Şelale Tüzel, 2014, Firm-level productivity, risk, and return, *Management Science* 60, 2073–2090.

Jones, Charles I., and Christopher Tonetti, 2020, Nonrivalry and the economics of data, *American Economic Review* 110, 2819–58.

Kogan, Leonid, Jun Li, and Harold H Zhang, 2023, Operating hedge and gross profitability premium, *The Journal of Finance* 78, 3387–3422.

Kogan, Leonid, and Dimitris Papanikolaou, 2013, Firm Characteristics and Stock Returns: The Role of Investment-Specific Shocks, *The Review of Financial Studies* 26, 2718–2759.

Kozlowski, Julian, Laura Veldkamp, and Venky Venkateswaran, 2018, The tail that keeps the riskless rate low.

Kuehn, Lars-alexander, Mikhail Simutin, and Jessie Jiaxu Wang, 2017, A labor capital asset pricing model, *The Journal of Finance* 72, 2131–2178.

Loualiche, Erik, et al., 2016, Asset pricing with entry and imperfect competition, *Journal of Finance, forthcoming* .

Ravn, Morten, Stephanie Schmitt-Grohé, and Martin Uribe, 2006, Deep habits, *The Review of Economic Studies* 73, 195–218.

van Binsbergen, Jules H, 2016, Good-specific habit formation and the cross-section of expected returns, *The Journal of Finance* 71, 1699–1732.

Veldkamp, Laura, and Cindy Chung, 2022, Data and the aggregate economy, *Journal of Economic Literature* .

Zhang, Lu, 2005, The value premium, *The Journal of Finance* 60, 67–103.

Appendix

A Model Appendix

A.1 Proof of Propositions

A.1.1 Proof of proposition 1

We normalize the aggregate price level P_t to one. The Lagrangian associated with the firm maximization problem (7) is

$$\begin{aligned}\mathcal{L} = \mathbb{E}_t \sum_{s=0}^{\infty} \Big[M_{t,t+s} \Big\{ & P_{i,t+s} Y_{i,t+s} - W^L_{t+s} L_{i,t+s} - I_{i,t+s} - W^N_{t+s} N_{i,t+s} - G^K_{i,t+s} K_{i,t+s} - G^N_{i,t+s} N_{i,t+s} \\ & + \lambda_{i,t+s} \Big(P^{-\eta}_{i,t+s} \mathcal{Y}_{t+s} + \Theta_{i,t+s} - Y_{i,t+s} \Big) + MC_{i,t+s} \Big(Z_{i,t+s} A_{t+s} F(K_{i,t+s}, L_{i,t+s}) - Y_{i,t+s} \Big) \\ & + q^K_{i,t+s}((1-\delta_K) K_{i,t+s} + I_{i,t+s} - K_{i,t+s+1}) + q^N_{i,t+s}((1-\delta_N) N_{i,t+s} + H_{i,t+s} - N_{i,t+s+1}) \Big\} \Big].\end{aligned}$$

We firstly take the first order condition with respect to the forecasting technology choice θ_{it}, this step closely follows Farboodi and Veldkamp (2021). By doing so, we can have the optimal technology $\hat{\theta}_{it} = \mathbb{E}_i[x_t|\mathcal{I}_{it}]$. Therefore, the expected forecasting technology is $\mathbb{E}[\Theta_{it}] = \bar{\Theta} - \mathbb{E}[(\mathbb{E}_i[x_t|\mathcal{I}_{it}] - x_t])^2]$. The second term is the conditional variance of x_t, which is Ω^{-1}_{it}. Therefore, the expected demand curve becomes Equation (9). We can rewrite the Lagrangian as

$$\begin{aligned}\mathcal{L} = \mathbb{E}_t \sum_{s=0}^{\infty} \Big[M_{t,t+s} \Big\{ & P_{i,t+s} Y_{i,t+s} - W^L_{t+s} L_{i,t+s} - I_{i,t+s} - W^N_{t+s} N_{i,t+s} - G^K_{i,t+s} K_{i,t+s} - G^N_{i,t+s} N_{i,t+s} \\ & + \lambda_{i,t+s} \Big(P^{-\eta}_{i,t+s} \mathcal{Y}_{t+s} + (\bar{\Theta} - \Omega^{-1}_{i,t+s}) - Y_{i,t+s} \Big) + MC_{i,t+s} \Big(Z_{i,t+s} A_{t+s} F(K_{i,t+s}, L_{i,t+s}) - Y_{i,t+s} \Big) \\ & + q^K_{i,t+s}((1-\delta_K) K_{i,t+s} + I_{i,t+s} - K_{i,t+s+1}) + q^N_{i,t+s}((1-\delta_N) N_{i,t+s} + H_{i,t+s} - N_{i,t+s+1}) \Big\} \Big].\end{aligned}$$

Take the first order condition with respect to data scientist N_{it}, price Y_{it},

$$q^N_{it} = \mathbb{E}_{it} \left[M_{t,t+1} \left\{ \nu_0 \nu_1 \sigma_s^{-2} \lambda_{i,t+1} \Omega^{-2}_{i,t+1} N^{\nu_1 - 1}_{i,t+1} Y^{1-\nu_1}_{i,t} - W^N_{t+1} - \frac{\partial G^N_{i,t+1}}{\partial N_{i,t+1}} N_{i,t} - G^N_{i,t+1} + q^N_{i,t+1}(1-\delta_N) \right\} \right], \tag{A1}$$

$$P_{it} = \lambda_{it} + MC_{it} - \mathbb{E}_t[\nu_0(1-\nu_1)\sigma_S^{-2} \lambda_{i,t+1} \Omega^{-2}_{i,t+1} N^{\nu_1}_{i,t+1} Y^{-\nu_1}_{it}], \tag{A2}$$

$$Y_{it} = \eta \lambda_{it} P^{-\eta-1}_{it} \mathcal{Y}_t. \tag{A3}$$

To simplify the notation, let's denote $\tilde{\lambda}_{i,t} = \mathbb{E}_t[\nu_0(1-\nu_1)\sigma_S^{-2}\lambda_{i,t+1}\Omega_{i,t+1}^{-2}N_{i,t+1}^{\nu_1}Y_{it}^{-\nu_1}]$. By combining the Equation (A1) and (A2), we will reach Proposition 1. □

A.1.2 Proof of proposition 3

By combining Equation (9) and (A3), we can derive the Proposition (3). □

B Additional Empirical Evidence

B.1 Impulse Responses of IT Related Investments

In Section 4.4, we have shown that the ratio of aggregate IT investment to physical investment is countercyclical. For robustness, we now present the IRFs of IT investment in response to aggregate TFP shocks, using the same recursive ordering but replacing the IT investment ratio with the log of IT investment. The results presented in Figure B.1 show that aggregate IT-related investment shows very little response to an aggregate TFP shock, hence the IT investment is acyclical.

C Data Appendix

C.1 Occupations related to data scientists

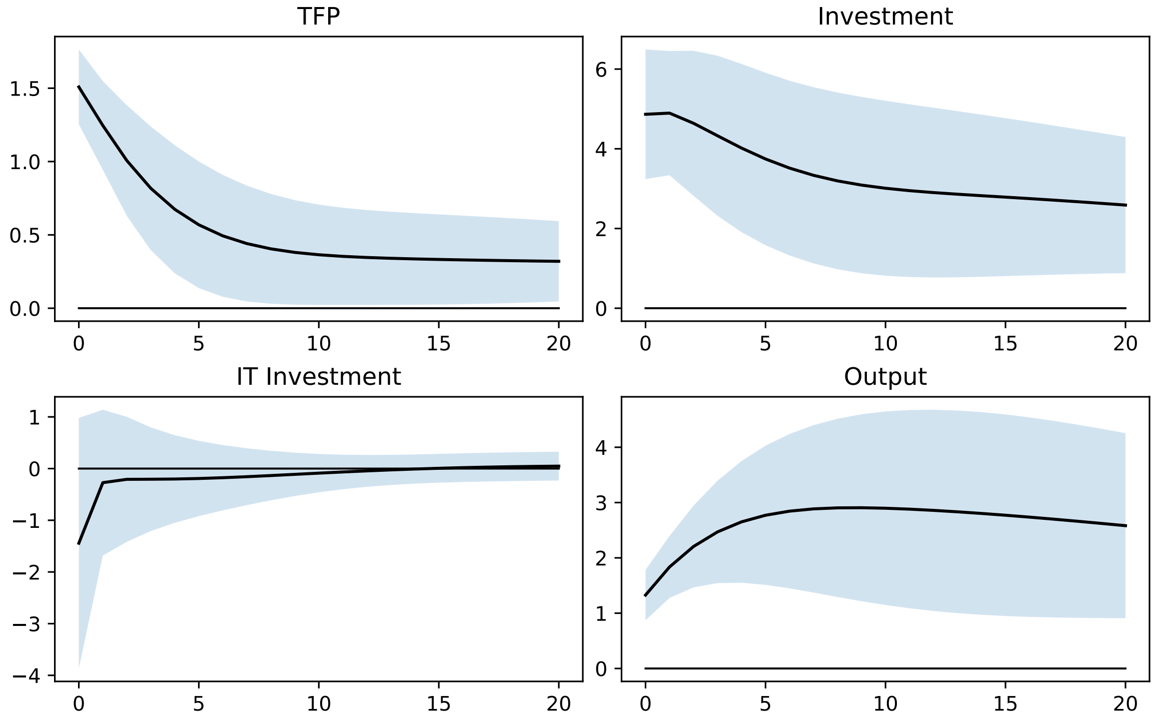
We use a broad set of definition for data scientists, the list of occupations includes data scientist, data analyst, data engineer, data administrator, economist, IT analyst, IT specialist, information specialist, information security, scientist, software developer, and software engineer. These job names are directly from Revelio's database. We keep these occupations related to data science because they play important supportive roles in data storage and analysis. A firm with more of these types of employees can do better in analyzing data.

C.2 Summary Statistics Across Industries

In Table C.1, we report summary statistics for data scientist ratios of firms in each industry according to the NAICS two-digit industry classifications. Some industries have more firms reporting to the hiring database, such as the Professional, Scientific, and Technical Services, Primary Metal Manufacturing, Construction, and Administrative and Support Service industries. There is a relatively large cross-industry variation in data scientist ratios. Specifically, the standard deviation of data scientist ratios ranges from 1.69% for the

Figure B.1: **Time-series Patterns of the IT Investment**

This figure plots the impulse responses of selected variables to an orthogonoalized one-standard deviation shock to the aggregate TFP shock. The shaded bands represent the 95% confidence intervals calculated from 1000 bootstrap replications. The sample period is 1955 to 2022, at an annual frequency.



Transportation and Warehousing industry to 10.82% for the Wood Product Manufacturing industry. Therefore, to make sure our results are not driven by any particular industry, we control for industry effects in our further analyses.

Table C.1: **Data Scientist Ratios across NACIS Two-digit Industries**

This table reports summary statistics of the industry-year observations of nonmissing data scientist ratios (%) across industries, including the pooled mean and median. Industries are based on NAICS two-digit industry classifications, excluding financial industries. The sample period is 2008-2021.

NAICS	**Industry Name**	**Obs**	**Mean**	**Median**	**Std**
21	Mining	1,612	3.76	3.07	3.54
22	Utilities	1,043	6.08	6.15	2.55
23	Construction	543	2.01	1.79	1.71
31	Food Manufacturing	1,270	3.52	2.73	5.31
32	Wood Product Manufacturing	4,796	10.41	6.09	10.82
33	Primary Metal Manufacturing	8,500	5.92	4.51	4.61
42	Wholesale Trade	1,079	3.89	3.16	3.28
44	Retail Trade	759	3.05	2.23	2.54
45	General Merchandise Retailers	913	2.93	2.54	1.73
48	Transportation and Warehousing	676	3.12	2.68	1.69
51	Information	3,643	11.65	10.64	7.68
53	Real Estate and Rental and Leasing	633	5.20	3.14	5.83
54	Professional, Scientific, and Technical Services	1,402	11.89	10.65	7.67
56	Administrative and Support Services	779	6.02	4.05	5.45
62	Health Care and Social Assistance	686	5.04	3.91	3.72
71	Arts, Entertainment, and Recreation	22	3.32	2.68	3.64
72	Accommodation and Food Services	696	1.30	0.79	1.83