Machine-Readable Data and Financial Analysts in Asset Management

Junli Zhao[*]

November 2022

Abstract

Machine-readable (clean and structured) data facilitate algorithm-driven investment decisions. Do financial analysts in asset management firms benefit from an increasing amount of machine-readable data? Exploiting an exogenous regulatory shock that makes corporate filings more machine-readable, I find that the performance of institutions with more financial analysts is impacted more positively when machine-readable data proliferate. In addition, these institutions increase their holdings on the stocks with more machine-readable data, both on the intensive and extensive margins. These results indicate that machine-readable data can benefit human analysts by increasing their information processing capacity.

Keywords: Information Technology; Skilled Labor; Information Acquisition

[*] Email: junli.zhao@city.ac.uk. Bayes Business School, City University of London. I thank Jean-Edouard Colliard, François Derrien, Olivier Dessaint, Thierry Foucault, Johan Hombert, Joel Peress, Christophe Spaenjers, and participants at HEC Paris, University of Vienna, and City University of London for helpful comments. All errors are my own.

1 Introduction

Computer algorithms are transforming the financial industry, with a potentially large impact on its labor force, in particular financial experts. On the one hand, these algorithms boost the productivity of financial experts by automating routine but complex tasks, enabling financial experts to generate more value. For example, Goldman Sachs' proprietary software system SecDB helps its financial experts evaluate the impact of the trades they propose.[1] On the other hand, algorithms have the potential to displace financial experts. For instance, robo-analysts are able to generate recommendations faster and better than human analysts (Coleman et al., 2020). As computers now manage about 35% of US public equities (vs. 24% for human asset managers),[2] understanding how computer algorithms may disrupt financial institutions and the finance labor market is important. As machine-readable (clean and well-structured) data are essential for algorithm-based analysis, identifying the relation between machine-readable data and financial experts helps evaluate the disruption.

Focusing on the production of information in asset management, this paper aims at investigating whether machine-readable data (or simply "data" hereafter) help financial experts generate more precise information (complementarity) or make them less essential in its production (substitution). Exploiting an exogenous regulatory shock that increased the amount of machine-readable data, I provide evidence consistent with complementarity.

The regulatory shock, the SEC's eXtensible Business Reporting Language (XBRL) mandate, requires firms to provide a machine-readable version of their corporate filings (10-K, 10-Q, etc.) using the XBRL format. Data items in the XBRL files are tagged with standard taxonomies. This feature makes it much easier for computers to extract information such as numbers in footnotes or numbers scattered in long paragraphs of texts, increasing the amount of data that are ready for large-scale computer-based analysis. Despite the different

[1] See, *Understanding SecDB: Goldman Sachs's Most Valued Trading Weapon*, the Wall Street Journal, Sept. 7, 2016

[2] The rest is owned by other investors, such as individuals and companies that are not asset managers. See, *March of the machines*, the Economist, June 11, 2019.

organization of data, the XBRL filings contain the same information as traditional text filings. The XBRL mandate was implemented through a three-year phase-in period from 2009 to 2011, during which large, medium and small firms complied successively. The staggered implementation provides a set-up to employ a (triple) difference-in-differences method.

To derive predictions that identify the relationship between data and financial experts, I consider a model in which information production depends on both data and financial experts. In the model, institutional investors trade one risk-free asset and several risky assets, conditioning their demands on market prices and private information about the payoffs of the risky assets. More data and financial experts improve their private information. The amount of data possessed by an institution is increasing in the number of computer scientists it employs and decreasing in the cost of data processing. I then consider comparative statics if the cost of data processing is lowered, as the XBRL mandate does.

Smaller processing cost increases the amount of machine-readable data, improving the investors' signal precision. How this improvement affects institutions with different numbers of financial experts depends on the relationship between the two inputs. If machine-readable data complement financial experts, i.e., data improve the productivity of financial experts, then when data grow, the signal precision of institutions with more financial experts increases relative to institutions with fewer financial experts. Thus institutions with more financial experts benefit more from the shock. Conversely, if the two inputs are substitutes, i.e., growth in data reduces the marginal value of financial experts in information production, the signal precision of institutions with more financial experts decreases relative to institutions with fewer financial experts.

More precise information decreases uncertainty, and thus affects performance and holdings. The model gives the following predictions on equilibrium performance and stock holdings, depending on whether the two inputs are complements or substitutes:

(i) The performance of institutions with more financial experts on the treated stock *does not decrease* (*decreases*) relative to that of institutions with fewer financial experts, given the same number of computer scientists;

(ii) The fraction of stock shares owned by institutions with more financial experts *does not decrease* (*decreases*) relative to that of institutions with fewer financial experts, given the same number of computer scientists;

(iii) Dispersion of holdings (standard deviation of holdings on the same stock across institutions) across finance-intensive institutions *does not increase* (*increases*) relative to that across base type (defined below) institutions.

An empirical investigation thus requires information on institutional investors' labor force, for which I use their foreign high-skilled labor (H-1B visa application) data. The numbers of IT- and finance-related workers, scaled by the institution's assets, provide a proxy for its labor inputs. The scaled labor input is labeled as high (low) if it falls in its distribution's top (bottom) tercile. With labels in both dimensions, I classify institutions into four types: (1) *base* type, i.e., low input of both types of workers; (2) *IT-intensive* type, high input of IT workers but low input of finance workers; (3) *finance-intensive* type, high input of finance workers but low input of IT workers; and (4) *bi-intensive* type, high input of both types of workers. The classification is verified by a manual check on some of the largest institutions in the data. Consistent with expectation, institutions well-known for their quantitative investment strategies are classified as IT-intensive while those classified as finance-intensive are more often associated with fundamental analysis. To test which relation holds empirically, I then compare differences across institutional investors in their performance and holding on the treated and control stocks, before and after the shock. In the analysis, I only consider two quarters before and after each event.

The predictions on performance are tested using the institutions' excess returns. For an institution, its excess return on stock j is computed as the product of its excess holding on asset j, relative to the market average holding, and the return of the stock. The annualized return of finance-intensive institutions increases by about 4 basis points on one treated stock relative to the base types after the XBRL shock, which is in line with financial experts benefiting from more machine-readable data. Similarly, the annualized return of bi-intensive type institutions increases by about 0.25 basis points on one treated stock relative to IT-intensive type institutions after the XBRL shock. As a byproduct, I find that the annualized return difference between the IT-intensive types and the base types decreases by about 3 basis points per stock after the shock. This result implies that IT-intensive institutions lose part of their informational advantage after the shock, which is consistent with that computer scientists help aggregate data and the information production function is concave in the amount of machine-readable data. Using changes in excess holdings as a proxy for trading, I also run the tests with returns from trading and obtain similar results.

For each stock in each quarter, I compute the fraction of total shares outstanding held by institutions of each type. Compared to base-type institutions, the fraction of total shares outstanding of the treated stocks owned by IT-intensive institutions decreases by 0.6 percentage points. This again is consistent with the hypothesis that the informational advantage of IT-intensive institutions is smaller after the shock. The shock has no significant impact on the fraction owned by base-type institutions and finance-intensive institutions at face value. One explanation for this result is that finance-intensive institutions are smaller than other institutions in the sample and thus changes in their fraction are mechanically smaller. To address this issue, I divide the fraction by the total assets of each type and obtain a measure of ownership per dollar for each stock and each type of institution. I find that, relative to the non-treated stocks, the ownership per dollar of treated stocks increases for base-type institutions and decreases for IT-intensive institutions after the implementation of the XBRL mandate. The increase in ownership per dollar is higher for finance-intensive institutions. These results are also consistent with complementarity.

The tests on the dispersion of excess holdings provide additional evidence for complementarity. Following the XBRL shock, the standard deviation of excess holding on the treated stocks decreases by 100 basis points among base-type institutions, compared to the non-treated stocks. In addition, compared to the base type, the effect is larger for the finance-intensive type (about 250 basis points).

As IT-intensive institutions are larger and finance-intensive institutions are smaller than the base-type institutions, one concern is that the results are driven by institution size. As a robustness check, I repeat the tests using a matched sample. I use coarsened exact matching on assets under management and turnover. Based on their values in the two variables, institutions are put into two-dimensional bins. Institutions within the same bin are matched. I construct a sample that only keeps institutions matched with the base type institutions. Differences in size are much smaller in this sample. Tests using this sample give results similar to the baseline analysis.

The rest of the paper is organized as follows. In Section 2, I lay out the theoretical framework and empirical predictions. Section 3 describes the shock and data in detail. The main results are presented in Section 4. Section 6 concludes.

Related literature. This paper contributes to a surging literature on the impact of IT, especially robots and artificial intelligence (AI), on labor. See, e.g., David (2015), Acemoglu and Restrepo (2018), Brynjolfsson et al. (2018) and Webb (2019). Based on regional data, Akerman et al. (2015) find that broadband internet improves the productivity of high-skilled workers. Several papers show how "machines" can complement human experts by reducing behavioral biases and providing new insights, in, e.g., earning forecast (van Binsbergen et al., 2020), real asset valuation (Aubry et al., 2019), or the selection of corporate directors (Erel et al., 2018).

Different from the previous literature, which mostly relies on simulation, this paper, using the actual investment behavior of institutional investors, provides evidence that high-skilled financial workers benefit from modern information technologies. Coleman et al. (2020) show that computers can generate investment recommendations faster and more accurately than human analysts. Grennan and Michaely (2020) find that sell-side analysts are more likely to shift their coverage or even leave the profession when their portfolio stocks are more exposed to AI analysis. Complementing their research, this paper suggests that access to better in-house technologies can help analysts mitigate the negative impact of AI on them.

Secondly, this paper is related to the fast-growing literature studying the effect of information technologies on financial institutions and markets. Modern information technologies have enlarged institutions' portfolio strategies (Abis, 2017). By easing access to public information and providing alternative datasets (Grennan and Michaely, 2019), modern IT technologies make price efficiency higher (Gao and Huang, 2020), investors disagree less (Chang et al., 2020), and corporations invest more (Goldstein et al., 2020). They also potentially contribute to the trend of rising price informativeness ((Bai et al., 2016), Farboodi et al. (2020)). Several studies investigate how improved technologies affect investors' information production. Farboodi and Veldkamp (2020) argue that technological progress in data processing can lead investors to rely more on data about others' demand than fundamentals but both types of data continue to be processed. Dugast and Foucault (2018) highlight the possibility that, with abundant data, investors might choose to rely more on raw signals than waiting for processed signals, which may reduce price informativeness. By looking into the information production function and different types of workers, this paper provides evidence that information technologies complement financial experts, especially by generating non-traditional information. Focusing on a different dichotomy, Abis and Veldkamp (2020) study how the productivity of data aggregation and data analysis skills in asset management evolved in the past decade. The results in this paper lend support to their assumption that the two skills are complements.

Thirdly, this paper is related to the large literature on information acquisition (e.g., Grossman and Stiglitz (1980), Verrecchia (1982); see Veldkamp (2011) for a survey). The precision of investors' signals plays an important role in this literature. Kacperczyk et al. (2016) develop a theory on how investors choose the precision of signals when they face an information capacity constraint. Van Nieuwerburgh and Veldkamp (2010) relate such information choice to mutual fund under-diversification. My paper puts a structure on signal precision that links precision and labor inputs and proposes a measure that can sort financial institutions by their information capacity.

Lastly, this paper is related to the literature on the impacts of the XBRL mandate and the empirical literature that exploited the H-1B visa program. Some studies suggest that the adoption of XBRL lowers information-aggregation costs. For example, using the XBRL Voluntary Filing Program before the mandate, Efendi et al. (2016) find that the market reaction is larger when XBRL reports are filed, indicating XBRL files have higher information value than HTML files. Other findings indicate that XBRL reduces event return volatility (Kim et al., 2012), analyst forecast dispersion (Liu et al., 2014), stock price synchronicity (Dong et al., 2014), increases the breadth of ownership (Kim et al., 2019), and quantitative footnote disclosure (Blankespoor, 2019). Bhattacharya et al. (2018) find that small institutions benefit more than large institutions from the mandate. This paper further shows that institutions with fewer resources in IT benefit more. Several studies have exploited H-1B data to study the effect of high-skilled labor on start-up success (Dimmock et al., 2019), the impact of skilled immigrant labor on innovation (Kerr and Lincoln, 2010), employment structure (Kerr et al., 2015), native wages and employment (Peri et al., 2015) and the cross-section of equity return (Sharifkhani, 2018).

2 Theoretical framework and empirical predictions

In this section, I first describe the theoretical framework that I rely on. Then I discuss its empirical predictions and tests. The set-up of the framework also lays down the terminologies that are used throughout this paper. A formal solution of the model is given in the appendix.

2.1 Framework set-up

I follow a noisy rational expectation equilibrium framework (see, e.g., Grossman and Stiglitz (1980), Admati (1985) and Kacperczyk et al. (2016)). There is one risky asset with payoff $f = \mu + z$, where $z \sim N(0, \sigma)$. σ is the prior variance. The supply of risky asset is $\bar{x} + x$, where $x \sim N(0, \sigma_x)$. There is a riskless asset with return r.

A unit measure of investors (asset management institutions), indexed by i, have exponential utility: $U_i = E[-exp(\rho W_i)]$, where $W_i = (W_{0i} - q_i p)r + q_i f$ is the end-of-period wealth of institution i and W_{0i} is its initial wealth. q_i is the number of shares of the risky asset that institution i chooses to hold.

Before trading, each institution receives a private signal about the assets' payoffs, the precision of which depends on the amount of machine-readable data it has and the number of financial analysts it employs. The signals η_{si} are independent and identically distributed across institutions, with $\eta_{si} = f + \epsilon_{si}$, where $\epsilon_{si} \sim N(0, K_i^{-1})$. K_i is the precision of institution i's signal. I assume $K_i = K(D_i, n_{fi}, n_{p_i})$. D_i is the amount of machine-readable data investor i has. n_{fi} is the number of finance analysts employed by institution i. n_{pi} is the number of computer specialists ("programmers") in the institution. The information production function is increasing in its arguments, i.e., $\frac{\partial K}{\partial D} > 0$, $\frac{\partial K}{\partial n_{fi}} > 0$ and $\frac{\partial K}{\partial n_{pi}} \geq 0$. In the rest of this paper, I suppress the third argument n_{pi} in K for brevity.

Computer specialists ("programmers") can also aggregate machine-readable data, for example, by using programs to parse traditional text corporate filings. For a given number of programmers, the amount of machine-readable data also depends on the difficulty or cost of processing data. To take the two observations into consideration, I assume that $D_i = D(c, n_{pi})$ with $\frac{\partial D}{\partial c} < 0$ and $\frac{\partial D}{\partial n_{pi}} > 0$. c is the cost of processing data of the risky asset.

The goal of this paper is to identify the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$: a positive sign means that machine-readable data and financial analysts are complements, whereas a negative sign indicates substitutability.

Institutions are characterized by their labor decomposition (n_{pi}, n_{fi}), which is exogenously given. To be closer to the empirical setting, I assume binary values for n_{pi} and n_{fi}: $n_{pi} \in \{\underline{n}_p, \overline{n}_p\}$ and $n_{fi} \in \{\underline{n}_f, \overline{n}_f\}$. The model thus features four types of institutions: base type with $(\underline{n}_p, \underline{n}_f)$, IT-intensive type with $(\overline{n}_p, \underline{n}_f)$, finance-intensive type with $(\underline{n}_p, \overline{n}_f)$ and bi-intensive type with $(\overline{n}_p, \overline{n}_f)$. The measure of type (n_p, n_f) institutions is denoted by $\mu(n_p, n_f)$

and constant throughout the model.

2.2 Hypothesis building

The equilibrium portfolio of institution i depends on the posterior variance $\hat{\sigma}_i$ and the expected payoff $\hat{\mu}_i - pr$. As signal precision is not directly observed in the data, I rely on their performance and stock holdings to derive empirical predictions. One commonly used measure of expected return is the excess return (see, e.g., Kacperczyk et al. (2016)). For an institution i, the excess return on the risky asset is defined as the unconditional expectation of the product of excess holding relative to the market, $q_i - \bar{q}$, and excess payoff of one unit of the asset, $f - pr$, namely,

$$E[R(n_{pi}, n_{fi})] = E[(q_i - \bar{q})(f - pr)]$$

Here $\bar{q}$ is the average holding on the risky asset across institutions and equals to $\bar{x} + x$ due to market clearing. The excess holding filters out holdings due to information conveyed by the market price and adjusts for risk. It is more sensitive to private information than the gross holdings q_i. Averaging across institutions with the same labor composition (n_p, n_f), a simple calculation gives that

$$E[R(n_p, n_f)] = (\rho \bar{x}^2 \bar{\sigma}^2 + \frac{1}{\rho} v)(K(D(c, n_p), n_f) - \bar{K}) \tag{1}$$

Here $\bar{K}$ is the average signal precision across institutions. $\bar{\sigma}$ is the average posterior variance of payoff. v measures the unconditional variance of the excess payoff $f - pr$. Given market variables ($x, \bar{\sigma}, \bar{K}$ and v), institutions' expected excess returns are increasing in their signal precision K.

To derive predictions that are informative about $\frac{\partial^2 K}{\partial D \partial n_f}$, I consider how excess returns change after an exogenous decrease in c. A smaller c implies more machine-readable data and thus better information for each institution. Depending on the labor composition and the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$, the impact ($\frac{\partial E[R(n_p,n_f)]}{\partial c}$) is heterogeneous across different types of institutions, as shown by the following results.

Proposition 1. *(i) If* $\frac{\partial^2 K}{\partial D \partial n_f} \leq 0$, $\frac{\partial E[R(n_p,\overline{n}_f)]}{\partial c} < \frac{\partial E[R(n_p,\underline{n}_f)]}{\partial c}$*. Substitution implies that institutions with more financial analysts benefit less from the shock than institutions with fewer financial analysts.*
(ii) $\frac{\partial E[R(n_p,\overline{n}_f)]}{\partial c} > \frac{\partial E[R(n_p,\underline{n}_f)]}{\partial c}$ *only if* $\frac{\partial^2 K}{\partial D \partial n_f} > 0$*. A situation where institutions with more financial analysts benefit more implies complementarity.*

These results are intuitive. The heterogeneous effects of a data increase come from two channels: (a) *Direct channel.* The direct effect on the precision of private signal depends on institutions' labor composition and the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$, i.e., the relation between the two inputs. If the two inputs are substitutes, an increase in the amount of machine-readable data decreases the marginal productivity of financial analysts. As a result, the signal precision of institutions with more financial analysts decreases relative to institutions with fewer financial analysts. If they are complements instead, more machine-readable data benefit institutions with more financial analysts even more, resulting larger informational advantage for institutions with more financial analysts. (b) *Market price channel.* As all institutions produce more precise information, the market price also incorporates more information. Institutions with fewer financial analysts produce less precise information than other institutions, holding the number of computer specialists constant, and they put more weight on information in the market price. Therefore, institutions with fewer financial analysts benefit more from a more informative market price. The impact of more data is more positive for them through this channel.

Note that part (i) is a sufficient condition while part (ii) gives a necessary condition. This asymmetry is because the two channels may have opposite impacts on institutions' informational advantage. In the case of substitution, both effects reduce the informational advantage of institutions with more financial analysts. In the case of complementarity, the market price channel works against the direct effect, which leads to the conclusion that a more positive impact on the performance of institutions with more financial analysts implies complementarity.

The exogenous shock on c also changes the unconditional holding of the institutions. For institutions with (n_p, n_f), their average unconditional holding on the risky asset is

$$E[q|(n_p, n_f)] = \hat{\sigma}(n_p, n_f)^{-1}\bar{\sigma}\bar{x} \tag{2}$$

where $\hat{\sigma}(n_p, n_f)$ is the posterior variance associated with type (n_p, n_f). The market clearing condition implies that unconditional ownership of stock by a type (n_p, n_f) investor is given by

$$\begin{aligned} E[Ownership|(n_p, n_f)] =& \frac{\mu(n_p, n_f)E[q|(n_p, n_f)]}{\bar{x}} \cdot \frac{1}{\mu(n_p, n_f)} \\ =& \hat{\sigma}(n_p, n_f)^{-1}\bar{\sigma} \end{aligned}$$

Similar to Proposition 1, the following result on stock ownership is immediate.

Proposition 2. *(i) If $\frac{\partial^2 K}{\partial D \partial n_f} \leq 0$, the fraction of the risky asset owned by institutions with more financial analysts decreases relative to the fraction owned by institutions with fewer financial analysts after the shock.*
(ii) The fraction of the risky asset owned by institutions with more financial analysts increases relative to the fraction owned by institutions with fewer financial analysts after the shock only if $\frac{\partial^2 K}{\partial D \partial n_f} > 0$.

The intuition behind this result is that higher signal precision decreases ex-ante uncertainty and hence increases unconditional ownership. The economic mechanism behind these two results is similar to the one on excess returns.

For holding dispersion, I use the standard deviation of excess holding for each type of institutions. Specifically, for type (n_p, n_f) institutions, their holding dispersion, $\delta_j(n_p, n_f)$, is defined as

$$\begin{aligned}\delta(n_p, n_f) \equiv& \sqrt{Var(q_i - \bar{q} \mid i \in (n_p, n_f))} \\ =& \frac{1}{\rho}\sqrt{v(K(D(c, n_p), n_f) - \bar{K})^2 + K(D(c, n_p), n_f)}\end{aligned} \tag{3}$$

$Var(\cdot)$ is the variance operator. $\delta(n_p, n_f)$ is the variance of excess holdings conditional on type (n_p, n_f). The following result shows that under certain mild conditions, the relation between D and n_f can be identified through a comparison between institutions with different numbers of financial analysts.

Proposition 3. *Assume σ_x and $K(D(c, \overline{n}_p), \overline{n}_f)$ are large enough, if $\frac{\partial^2 K}{\partial D \partial n_f} \leq 0$, then $\delta(\underline{n}_p, \overline{n}_f)$ decreases less than $\delta(\underline{n}_p, \underline{n}_f)$ when c decreases. If $\delta(\underline{n}_p, \overline{n}_f)$ decreases no less than $\delta(\underline{n}_p, \underline{n}_f)$, then $\frac{\partial^2 K}{\partial D \partial n_f} > 0$.*

Holding dispersion depends both on the distance to the average precision $(K(D(c, n_p), n_f) - \bar{K})^2)$ and the institutions' own precision $K(D(c, n_p), n_f)$. Note that as institutions are collectively more informed, the unconditional variance v decreases. When the supply of the asset is noisy enough (σ_x is large enough), the first term in the square root dominates, and $\delta(\underline{n}_p, \underline{n}_f)$ decreases. In this case, how $\delta(\underline{n}_p, \overline{n}_f)$ changes relative to the dispersion of holdings across the base type depends on the relation between D and n_f.

The conditions of this result can be easily verified empirically by examining the change in $\delta(\underline{n}_p, \underline{n}_f)$ following the decline of c. A decrease in $\delta(\underline{n}_p, \underline{n}_f)$ confirms that in equation (3) the distance term $(K(D(c, n_p), n_f) - \bar{K})^2)$ dominates as the precision term (K) implies an increase.

2.3 Empirical predictions

The previous results suggest that we can infer the relationship between machine-readable data and financial analysts by exploiting a shock that decreases the cost of data processing on a subset of stocks, using the following regression,

$$\begin{aligned} y_{ijt} =& \gamma_\theta Type_{i,\theta} \times Treated_j \times Post_t + \beta\ Treated_j \times Post_t + Type_{i,\theta} \times Treated_j \\ &+ Type_{i,\theta} \times Post_t + Treated_j + Type_{i,\theta} + Post_t + Controls \end{aligned} \tag{4}$$

$Type_{i,\theta}$ is an indicator for institution i on whether i belongs to type θ (IT-, finance-, or bi-intensive type). $Treated_j$ is an indicator for stock j on whether the cost of processing its data is affected by the shock. $Post_t$ indicates whether the shock has taken place. γ_θ captures the differential impacts between $Type_\theta$ and the base type institutions. β captures the impact of the shock on treated stocks relative to non-treated stocks for the base-type institutions. Using excess returns as the dependent variable, Proposition 1 indicates the following tests.

Test 1. *In the case of substitution, $\gamma^r_{finance} < 0$: the impact of the shock on the performance of base-type institutions is more positive than that on the performance of finance-intensive institutions.*

Test 2. *$\gamma^r_{finance} \geq 0$ implies complementarity, i.e., if the impact of the shock on the performance of finance-intensive institutions is more positive than that on the performance of base-type institutions, then machine-readable data complement financial experts.*

Tests 1 and 2 also hold when using stock holdings as the dependent variable, in which case a positive $\gamma^s_{finance}$ from regression (4) implies complementarity while a negative γ^s_θ is consistent with they being substitutes.

Running regression (4) using holding dispersion as the dependent variable gives γ^δ_θ. Proposition 3 provides the following predictions.

Test 3. *$\gamma^\delta_{finance} \leq 0$ implies complementarity. In the case of substitution, $\gamma^\delta_{finance} > 0$: if holding dispersion decreases more for finance-intensive investors than for the base type investors, the two inputs are complements; if the two inputs are substitutes, holding dispersion decreases less for finance-intensive investors.*

Test 3 provides another way to determine how machine-readable data interact with financial experts. Next, I carry out these tests using the SEC's XBRL mandate as a shock on the cost of data processing.

3 Empirical setup

3.1 The XBRL mandate and sample construction

I use the implementation of the SEC's XBRL mandate in 2009 as a shock that reduces the cost of data processing. XBRL (eXtensible Business Reporting Language) is a programming language that facilitates the communication of large volumes of business information using standard taxonomies and tagging. When preparing a financial statement in XBRL format, companies identify and tag each element in the statement with the standard taxonomy developed by the SEC. The tags are linked to their descriptive information, such as name, year, units, detailed definition, and also their relationship with other items. These features enable the users to easily locate the items and related information they are interested in. In comparison, gathering information from the static files (HTML or plain text) in the EDGAR system is time-consuming and costly. Although HTML files are also organized using tags, those tags are mostly location-based, carry little information about the content, and can vary across files. To get a sense of the difficulty, think about how to extract numbers from a footnote, which are usually scattered in text. If an analysis requires such information, the users, without the XBRL files, must either search for the information manually or develop sophisticated textual analysis programs that are based on location and context to extract information. Both methods require significant efforts and costs and are prone to error.

The identification strategy of this paper exploits the staggered implementation of the XBRL mandate: firms with a public float larger than 700 million USD are required to comply from June 15, 2009; firms with a public float between 50 million USD and 700 million USD must report in the XBRL format from June 15, 2010. For the rest, the mandate came into effect in June 2011. In principle, firms may voluntarily choose to disclose in the XBRL format before they are obliged to. As long as the early compliance is not meant to benefit a special type of institutions, it does not pose a threat to the identification strategy of this paper.

In each quarter, a stock's XBRL status is inferred from its file format in the SEC's EDGAR system. It is labeled as an XBRL stock (treated) if it has filed 10-K, 10-Q or 8-K in the XBRL format in that quarter. Figure 2 plots the time series of the number of XBRL stocks. As shown in the graph, there are three major dates when firms comply with the XBRL mandate: 2009 Q3, 2010 Q3 and 2011 Q3. In the analysis, I focus only on two quarters before and after each event (one cohort). I then stack observations from all cohorts together and align them along the period relative to the XBRL event. This construction avoids using the same observation as both treated and control. The sample period is thus from 2009 Q1 to 2011 Q4.

[Insert Figure 2]

3.2 Classification of institutions

Since data on institutional investors' entire labor force is not available, I use their foreign skilled labor as a proxy. Information on foreign skilled labor is obtained through the Labor Condition Application (LCA) data from the U.S. Department of Labor. This dataset contains each employee's job title, brief job description, and proposed contract duration. Such information allows me to construct a panel of institutions' desired number of foreign high-skilled workers in both IT and finance-related positions at each point in time. The LCA is a prerequisite for the H-1B visa, which is a temporary program that permits foreign skilled-workers in specialty-occupations to work in the U.S. These occupations require theoretical and practical application of highly specialized knowledge like engineering or accounting and attainment of a bachelor's or higher degree. An H-1B visa permits the holder to work in the U.S. for three years and can be renewed for a maximum of six years. The application for the H-1B visa has to be sponsored by an employer.

I classify jobs into IT- and finance-related positions based on job codes and job titles. For observations after mid-2009, I only use the Standard Occupational Classification (SOC) from the Department of Labor. IT jobs are positions with the SOC code that starts with 15-11 (Computer Occupations) or 11-3021 (Computer and Information Systems Managers). Finance jobs are positions with the SOC code that starts with 13-20 (Financial Specialists) or 11-3031 (Financial Managers). For observations before mid-2009, the job code classification is based on the Occupational Title (OT) codes from U.S. Citizenship and Immigration Services. For IT-related jobs, I include jobs with the OT code starting with 03 (Computer-Related Occupations), and 199 (Miscellaneous Professional, Technical, and Managerial Occupations) if the job title mentions one of the following words: developer, software, system, program, and information. For finance-related jobs, I include jobs for which the OT code starts with 50 (Occupations in Economics), 186 (Finance, Insurance, and Real Estate Managers and Officials), and 199 (Miscellaneous Professional, Technical, and Managerial Occupations) if the job tile mentions one of the following words: analyst, research, financial and investment. Using the information on contract duration and assuming no separation, I calculate the total number of both types of jobs for each institution in each quarter. As these numbers proxy cumulative hiring over the past three years (the maximum and the most frequent contract duration), they are more likely to reflect the current labor composition than only just recent recruitment.

An underlying assumption of this measure is that institutional investors on average do not particularly prefer native skilled workers for either IT or finance positions. With this assumption, the distribution of foreign skilled workers can proxy for the distribution of skilled workers in the entire workforce. If it is violated, for example, if native financial analysts are preferred, then many finance-intensive institutions would be classified as the base type. This goes against finding a significant difference between base-type institutions and the other types. There may be another concern that the measure based on the LCA data can be problematic since the visa may not be granted. Most foreign graduates benefit from the Optional Practical

Training (OPT) program, which allows them to work for at least one year without holding other visas. It is a common practice for firms to hire new employees relying on the OPT program and apply for the H-1B visa in advance before the OPT program expires. Even if the visa is not granted, the firm then may search for another employee to refill the position. Therefore the LCA data likely reflects the firm's desired number of positions.

To classify institutions into the four types, I merge the previous panel with institutional investor data from the Thomson Reuters 13-F dataset, using the names of the institutions. I keep only institutions with at least one H-1B visa application between 2007 and 2013. I exclude banks and insurance companies (Thomson Reuters' type codes 1 and 2). For each cohort, I then sort institutions based on their IT and finance intensity, which is the ratio between their number of IT or finance positions and their assets under management (AUM), both measured one quarter before the event. If an institution falls in the top (bottom) tercile of IT intensity distribution and in the bottom (top) tercile of finance intensity distribution, I classify it as an IT-intensive (finance-intensive) type. A base-type institution falls in the bottom tercile in both dimensions while a bi-intensive type assumes top terciles. The summary statistics of jobs in an institution are given in the first two rows of Table 1. The summary statistics of investor size by their type are described in row 3 through row 6. In the sample, there are about 420 institutions. IT-intensive and base-type institutions are relatively larger than finance-intensive and bi-intensive institutions.

[Insert Table 1]

One way to verify that the classification is sensible is to check whether the institutions' type roughly corresponds to their investment philosophy. An institution can probably rely more on quantitative methods if it has many computer specialists and more on discretionary or fundamental methods if its team is mainly composed of financial analysts. Table 2 reports the twenty largest institutions for IT-intensive, finance-intensive, and bi-intensive types. Consistent with this intuition, the measure classifies institutions which are well-known for their quantitative approach, such as D. E. Shaw & Co., Renaissance Technologies, or Two

Sigma Investments, as IT-intensive institutions. Institutions labeled as finance-intensive seem to rely more on fundamental analysis. For example, Tremblant Capital states on its website that its managers “conduct deep fundamental research to uncover investments that are trading at a material dislocation from fair value.” Sandler Capital Management believes that “in-depth fundamental research and deep industry knowledge are the primary contributing factors to successful investing.” Sirios Capital Management identifies itself as “a fundamentally-driven investment firm ... and its investment process is driven by fundamental research on a company-by-company basis.” Many large asset management firms, such as Merrill Lynch, UBS Securities and Bridgewater Associates are classified as the bi-intensive type. These facts lend confidence that the measures can capture an institution’s advantage in each of the two dimensions.

Table 3 provides further comparison on the different types of institutions. In Table 3, I compare each type of institutions to the base type using the following regression:

$$y_{it} = \alpha + \sum_{\theta} \beta_\theta Type_{i\theta t} + \alpha_t + Controls_{it}$$

The dependent variable in columns (1), (2), and (3) are institutional turnover, average market capitalization of portfolio stocks (weighted by portfolio weight), and log number of portfolio stocks, respectively. The result in column (1) shows that comparing to base-type institutions, IT-intensive institutions have higher turnover whereas finance-intensive institutions have lower turnover. This is also consistent with the intuition that IT-intensive institutions may exploit high frequency information, e.g. order flow, more often and have a shorter investment horizon. By contrast, finance-intensive institutions may rely more on fundamental approaches and have a longer investment horizon. Column (2) shows that on average finance-intensive institutions hold stocks with a smaller market capitalization. Column (3) suggests that these institutions hold similar number of stocks once their size is controlled.

3.3 Other Variables

Now I turn to the construction of other variables such as excess returns, stock ownership, dispersion in holding, and control variables.

Following Kacperczyk et al. (2016), I compute the excess return of investor i on stock j at time t as the product of its excess holding, $w_{ijt} - \bar{w}_{ijt}$, and the return on the stock, r_{jt},

$$R_{ijt}^{holding} = (w_{ijt} - \bar{w}_{jt})r_{jt}$$

where w_{ijt} is the weight of stock j in the portfolio of institution i in quarter t. $\bar{w}_{jt}$ is the average of w_{ijt} across all the institutions in the sample. r_{jt} is stock j's cumulative return in quarter t. Stock return data is from the CRSP. In theory, each institution should hold almost every stock to gain from diversification, few of them do in reality, either due to fixed costs or information capacity constraints. In the analysis, I only consider non-zero weights. The returns are measured in basis points.

Given that institutions may have different investment horizons, they may not adjust their positions on a given stock simultaneously. This feature makes the tests based on excess returns less powerful. One measure that may alleviate this concern is excess returns conditional on trading, which I construct as follows,

$$R_{ijt}^{Trading} = [(w_{ij,t} - \bar{w}_{ij,t}) - (w_{ij,t-1} - \bar{w}_{ij,t-1})]r_{jt}$$

where $(w_{ij,t} - \bar{w}_{ij,t}) - (w_{ij,t-1} - \bar{w}_{ij,t-1})$ is the change in investor i's excess holding on stock j between $t-1$ and t, which proxies its trading. To avoid mechanical change due to movements in sample averages or trading in other assets, I only consider observations with a change in the number of shares held.

I construct two measures of stock ownership. The first one is the fraction of total shares outstanding owned by a type of institution. Formally,

$$Fraction_{j\theta t} = \frac{\sum_i (Shares_{ijt} \cdot Type_{i,\theta})}{Total\ Shares_{jt}}$$

$Share_{ijt}$ is the number of shares of stock j held by institution i in quarter t. $Type_{i,\theta}$ indicates whether institution i is of type θ. Because types with larger institutions mechanically have higher ownership fractions, their changes may also be mechanically larger. To address this concern, I consider a second measure, $FracScaled$, which is the fraction scaled by the total assets of each type.

$$FracScaled_{j\theta t} = \frac{Fraction_{j\theta t}}{\sum_i (Assets_{it} \cdot Type_{i,\theta})} \times 10^6$$

The measure is multiplied by the constant 10^6 simply to avoid too many leading zeros in the estimates.

The holding dispersion $\delta_{j\theta t}$ is calculated as the standard deviation of excess holding, $w_{ijt} - \bar{w}_{ijt}$, for each stock j in each quarter t conditional on type θ institutions. The dispersion is measured in percentage points.

To control for an institution's other characteristics, I include their turnover and assets under management. Following Ben-David et al. (2010), an institution's turnover is the ratio between its total trading value for a given quarter and its assets under management.

4 Empirical Results

4.1 Substitutes or Complements ?

To investigate the relation between machine-readable data and financial analysts, I conduct the tests outlined in section 2 with the following regressions.

$$
\begin{aligned}
y =& \sum_{\theta} \gamma^r_\theta \; Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^r XBRL_{jc} \times Post_q + Controls_{ijqc} \\
& + \text{Stock-Type-Cohort FE} + \text{Type-Period-Cohort FE} \; (+\text{Institution FE}) \qquad (5)
\end{aligned}
$$

The dependent variables are excess returns (R_{ijqc}), stock ownership ($Fraction_{j\theta qc}$ and $FracScaled_{j\theta qc}$) or holding dispersion ($\delta_{j\theta qc}$). Period q is measured as the number of quarters relative to the event time in each cohort. Indicating the period q and cohort c together is equivalent to indicating the calendar quarter t. $Type_{i\theta c}$ is the type indicator for institution i in cohort c. The sum is over all the types except the base type. $XBRL_{jc}$ is equal to one if stock j complies to the XBRL mandate in cohort c and equals zero otherwise. $XBRL_{jc}$, $Type_{i\theta c}$ and $Type_{i\theta c} \times XBRL_{jc}$ are absorbed by Stock-Type-Cohort fixed effects. $Post_q$ and $Type_{i\theta c} \times Post_q$ are absorbed by Type-Period-Cohort fixed effects. I also report results that control for Stock-Period-Cohort fixed effects, which absorb $XBRL_{jc} \times Post_q$. In the regression on the excess returns, I control for institution fixed effects. Control variables such as assets under management, institution turnover and stock market values are included. In the regressions on stock ownership and holding dispersion, I control for stock market capitalization, shares owned by institutional investors, book-to-market ratio and leverage ratio. I also control for the average asset under management of the institutions that owns the stock for each type. γ_θ measures the difference between type θ institutions and base-type institutions on the XBRL stocks, before and after the shock, relative to the unaffected stocks.

4.1.1 Evidence on Excess Returns

The results of regression (5) are reported in Table 4. Columns (1) and (2) report the results using excess holding returns as the dependent variable. In column (1), the excess returns on the treated stocks increase more for finance-intensive institutions. Comparing to the control stocks, the annualized excess return on one treated stock is about 0.8 basis points higher for finance-intensive institutions than for base-type institutions. According to tests 1 and 2, this result implies complementarity and rejects substitution. The effect is also economically significant. Given that finance-intensive institutions hold on average about 70 stocks in the sample, this effect would translate into a 2.24 percentage points increase in their annual performance if all of the stocks are treated and the effect is homogeneous. The point estimate for Bi-intensive institutions is higher for IT-intensive institutions, also consistent with complementarity instead of substitution, even though the difference is not statistically significant.

IT-intensive institutions generate lower excess returns on the treated stocks by about 0.6 basis point than base-type institutions. For an IT-intensive institution that holds the average number of stocks (148), this decrease implies that the annualized performance gap decreases by 3.5 percentage point if all stocks are treated. This negative coefficient implies that IT-intensive institutions lose part of their informational advantage on the treated stocks after the shock. One possible explanation for this result is that IT-intensive institutions were able to process more data on the stocks since they have more computer specialists, which gives them an edge over institutions with fewer computer specialists. The mandate decreased the cost of data processing so that their advantage coming from more machine-readable data is diminished. In column (2), I control for stock-period-cohort fixed effects, which absorb stock fixed effects and $XBRL_{jc} \times Post_q$. The results are qualitatively similar.

The previous tests implicitly assume that institutions adjust their positions on each stock each quarter. However, this assumption may be less true if institutions have long investment horizons, which is probable especially for finance-intensive institutions. Although investment horizons are not directly observed in the data, this concern can be partly alleviated by conditioning the comparison on observations in which the institutions adjust their positions. Comparisons on these conditional observations can give more powerful tests. With is logic, columns (3) and (4) of Table 4 report the results with trading returns. The results are similar to the results based on holding returns and indeed suggest complementarity at a stronger statistical confidence level.

[Insert Table 4]

[Insert Figure 3]

Figure 3 and 4 plot the estimates of $Type \times XBRL \times Dummy_q$ for the holding and trading excess returns, respectively. These estimates capture the time evolution of the performance differences between finance-intensive (IT-intensive) type and base-type institutions on the treated stocks compared to the non-treated stocks. We expect the events to have positive impacts on the performance difference for the finance-intensive institutions. As shown in the graph, before the implementation, there is no significant difference, suggesting that the parallel trend assumption is not violated. The two performance gaps diverge after the treatment: the impact of lower cost of data processing is more positive (negative) for the finance-intensive (IT-intensive) institutions. In these tests, I compare quarterly performance of these institutions regardless of their investment horizons. The fact that the magnitudes of the impacts increase over time suggests that comparisons at longer horizons should generate qualitatively similar results.

[Insert Figure 4]

4.1.2 Evidence on Stock Ownership

Table 5 reports the results of tests based on stock ownership.

[Insert Table 5]

The dependent variable in the first two columns is the fraction of total share outstanding owned by each type of institutions. In column (1), compared to base-type institutions, the ownership on the treated stocks by IT-intensive and bi-intensive investors decreases by about 0.007 and 0.002 percentage point, or 23% and 6% of the sample average, respectively, after the implementation of the XBRL mandate. In column (2), the result is similar even after controlling for stock-period-cohort fixed effects. In both columns, the coefficient on the finance-intensive ins is negative though insignificant. One potential explanation for this finding is that finance-intensive institutions have less assets than other institutions in the sample. Even if they increase their holdings of the treated stocks more than base-type institutions, their ownership of treated stocks might not change significantly. To mitigate the effect of institution size, in column (3) and column (4), I divide the fraction of ownership by the total assets of each type. In column (3), scaled ownership by base-type institutions increases on the treated stocks comparing to the control group after the implementation of XBRL. In column (3) and column (4), scaled ownership by finance-intensive investors increased further than base-type institutions. For IT-intensive and bi-intensive institutions, the results are similar to columns (1) and (2). These results are consistent with our previous findings. They indicate the machine-readable data complement the financial analysts in information production. Lower cost of data processing erodes the informational advantage of IT-intensive institutions.

Figure 5 plots the time evolution of the differences in scaled ownership between the IT-intensive or finance-intensive institutions and base type institutions on the treated stocks, relative to the control stocks. Similar to the results in excess returns, the magnitudes of the impacts increase over time.

[Insert Figure 5]

4.1.3 Evidence on Stock Holding Dispersion

The results on dispersion of excess holdings are shown in Table 6. In column (1), the negative coefficient before $XBRL \times Post$, even though not significant, shows that the XBRL mandate decreases dispersion on the treated stocks among base type institutions, which validates our assumptions in Proposition 3 and test 3. The coefficient before the interaction with the finance-intensive indicator is negative but not statistically significant. One reason for this insignificant result may be that stock-type-cohort fixed effects are included in the regression. These fixed effects make sure that the comparison is among the same stock-type pair in each cohort but it also restricts the comparison to at most four observations: two before and two after the event, dramatically reduces the statistical power. In columns (3) and (4), I control for treated-type-cohort fixed effects, which is still reasonably conservative. After the implementation of the XBRL mandate, the dispersion on the treated stocks among base-type institutions decreased by 0.1 percentage point (p-value <0.1), or 20% of the sample mean. For finance-intensive institutions, the dispersion decreased further by 0.15 percentage point (p-value<0.05). According to test 3, this result again implies complementarity between machine-readable data and financial analysts.

[Insert Table 6]

Overall, the evidence on excess returns, stock ownership, and dispersion of excess holdings are more consistent with machine-readable complement financial analysts in information production. I now provide additional results using a matched sample.

4.2 Robustness Checks

One concern for the results is that the IT-intensive institutions are much larger in terms of assets (median=1993 million) than base-type institutions (median=545 million) while the finance-intensive institutions are smaller (median=81 million). The results may be driven by investor characteristics related to their size. It is not surprising that IT intensity is correlated with size. Given fixed set-up costs of an IT system, larger institutions may find it more economical to have more IT resources than small institutions. However, the comparison between the bi-intensive type and base type suggests that the size effect is not likely to explain the results. In the sample, the by-intensive institutions (median=109 million) are also smaller than base-type institutions. Yet, the estimation results are very different from the results on finance-intensive institutions.

To make the base type and other types more comparable in terms of size and investment horizon, I match the base type with other types using the Coarsened Exact Matching method in each cohort, based on their assets under management and turnover, both measured one quarter before an event. Table 7 gives the summary statistics of size by their type in the matched sample. Base$_i t$, Base$_f in$ and Base$_b i$ refers to base-type institutions matched with IT-intensive, finance-intensive and bi-intensive institutions, respectively.

I run similar regressions as in (5) on the three types, IT-intensive, finance-intensive and bi-intensive separately, relative to base-type institutions. The results are reported in Table 8 and Table 9. The results are not qualitatively changed and also consistent with machine-readable data and financial analysts being complements.

[Insert Table 8]

[Insert Table 9]

5 Evidence on the extensive margin

The previous empirical design focuses on stocks held by the institutions both before and after the treatment. In this section, I explore how portfolio holdings are impacted by the data shock. This investigation also sheds light on how machine-readable data boosts the productivity of human analysts.

Researchers have found that limited attention or information capacity may hinders investors from processing information at a large scale and thus from holding a very diversified portfolio. One way in which the machine-readable data help human analysts is easing and accelerating information processing, hence effectively increasing their information capacity. A direct implication is that following the data shock, finance-intensive institutions add more treated stocks into their portfolio. This is confirmed in Figure 1. In this figure, I plot the average number of distinct stocks held by each type of institutions, separately for the control stocks and the treated stocks. As shown in the graph, Finance-intensive institutions

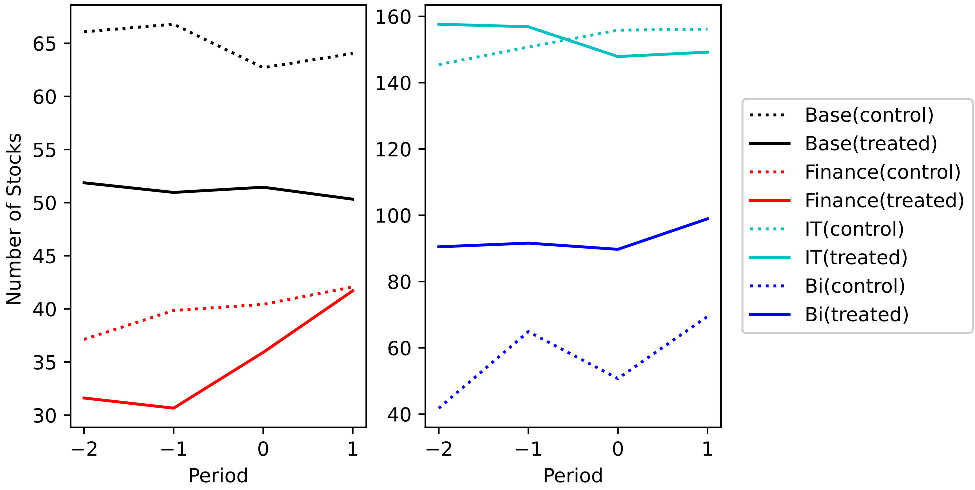

Figure 1: Average number of distinct treated/control stocks held across a given type of institutions.

expanded their holdings on the treated stocks from around 30 to 40, a 30% increase, upon more machine-readable data become available. In comparison, there is only a slight increase in the number of control stocks. Such dramatic contrast is not observed for other types of institutions. For IT-intensive institutions, an opposite pattern is observed. They divest some of the treated stocks and expanded their investment frontier in the control stocks possibly because they lose part of information advantage in the treated stocks. For the base-type, the average number of the control stocks decreases slightly while that of the treated stocks remains relatively stable. Interestingly, for the bi-intensive type, their portfolio expanded both in the treated and the control group. This pattern does not contradict with our conjecture. Indeed, one possible explanation is that when the shock increases their information processing capacity, they may expand their investment universe across both groups as they rely less on the shock to obtain and process data, similar to the IT-intensive type, and need not only focus on the treated stocks.

A natural question is how the portfolio returns of these institutions are affected. Proceeding similarly as before, I compare the excess portfolio returns of different types of institutions separately for the treated and the control stocks. Table 10 shows the results. Different from evidence at the intensive margin, the impacts of the data shock on the differences in excess portfolio returns are small and insignificant with an exception for IT-intensive institutions, whose excess returns on the treated group decreased significantly. Given our findings in the intensive, the null results arise because relatively poor performance on the newly added treated stocks erodes part of performance gain on the existing treated stocks. The null results imply that the overall expected performance of finance-intensive institutions do not gain much from an increase in machine-readable data. The benefit lies in enhanced portfolio diversification, which may lower the risk of their portfolios and deliver utility gain.

[Insert Table 10]

6 Conclusion

The rapid progress of information technologies brings deep changes to the financial industry. The ability of computers to process massive amounts of numbers, images, and natural languages demonstrates their potential not only for enhancing productivity but also replacing human labor. Using information on highly skilled labor in the finance industry, this paper suggests that, at least until recently, financial institutions exploited modern technologies mostly to assist their financial analysts. The proliferation of data complement human financial analysts in information production, despite the concern that their value is eroded by machines.

Understanding this complementarity helps job seekers or employers to make better career or recruitment decisions, and also helps regulators and policymakers to evaluate the impact of new technologies on the finance industry. Whether a similar complementarity can be found in other areas of the financial industry and how the relation might be changed by new technologies remains an open question. Using more recent changes or technology shocks, the methodology used in this paper can also shed light on these issues. As shown in this paper, the benefits of data availability is not equally shared by different market participants. Regulators and researchers can apply this method to identify winners and losers from other regulations or technologies.

References

Abis, S. (2017). Man vs. machine: Quantitative and discretionary equity management. *Unpublished working paper. Columbia Business School.*

Abis, S. and Veldkamp, L. (2020). The changing economics of knowledge production. *Available at SSRN 3570130.*

Acemoglu, D. and Restrepo, P. (2018). Artificial intelligence, automation and work. Technical report, National Bureau of Economic Research.

Admati, A. R. (1985). A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica: Journal of the Econometric Society*, pages 629–657.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The skill complementarity of broadband internet. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Aubry, M., Kräussl, R., Manso, G., and Spaenjers, C. (2019). Machine learning, human experts, and the valuation of real assets. *HEC Paris Research Paper No. FIN-2019-1332.*

Bai, J., Philippon, T., and Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654.

Ben-David, I., Glushkov, D., and Moussawi, R. (2010). Do arbitrageurs really avoid high idiosyncratic risk stocks? *Available at SSRN 1572955.*

Bhattacharya, N., Cho, Y. J., and Kim, J. B. (2018). Leveling the playing field between large and small institutions: evidence from the sec's xbrl mandate. *The Accounting Review*, 93(5):51–71.

Blankespoor, E. (2019). The impact of information processing costs on firm disclosure choice: Evidence from the xbrl mandate. *Journal of Accounting Research*, 57(4):919–967.

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108, pages 43–47.

Chang, Y.-C., Hsiao, P.-J., Ljungqvist, A., and Tseng, K. (2020). Testing disagreement models.

Coleman, B., Merkley, K. J., and Pacelli, J. (2020). Man versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *Available at SSRN 3514879*.

David, H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30.

Dimmock, S. G., Huang, J., and Weisbenner, S. J. (2019). Give me your tired, your poor, your high-skilled labor: H-1b lottery outcomes and entrepreneurial success. Technical report, National Bureau of Economic Research.

Dong, Y., Li, O. Z., Lin, Y., and Ni, C. (2014). Does information processing cost affect firm-specific information acquisition?-evidence from xbrl adoption. *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming*.

Dugast, J. and Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial Economics*, 130(2):367–391.

Efendi, J., Park, J. D., and Subramaniam, C. (2016). Does the xbrl reporting format provide incremental information value? a study using xbrl disclosures during the voluntary filing program. *Abacus*, 52(2):259–285.

Erel, I., Stern, L. H., Tan, C., and Weisbach, M. S. (2018). Selecting directors using machine learning. Technical report, National Bureau of Economic Research.

Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2020). Where has all the data gone? Working Paper 26927, National Bureau of Economic Research.

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

Gao, M. and Huang, J. (2020). Informing the market: The effect of modern information technologies on information production. *The Review of Financial Studies*, 33(4):1367–1411.

Goldstein, I., Yang, S., and Zuo, L. (2020). The real effects of modern information technologies. Technical report, National Bureau of Economic Research.

Grennan, J. and Michaely, R. (2019). Fintechs and the market for financial analysis. *Michael J. Brennan Irish Finance Working Paper Series Research Paper*, (18-11):19–10.

Grennan, J. and Michaely, R. (2020). Artificial intelligence and high-skilled work: Evidence from analysts. *Available at SSRN*.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, 70(3):393–408.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kerr, S. P., Kerr, W. R., and Lincoln, W. F. (2015). Skilled immigration and the employment structures of us firms. *Journal of Labor Economics*, 33(S1):S147–S186.

Kerr, W. R. and Lincoln, W. F. (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics*, 28(3):473–508.

Kim, J.-B., Li, B., and Liu, Z. (2019). Information-processing costs and breadth of ownership. *Contemporary Accounting Research*, 36(4):2408–2436.

Kim, J. W., Lim, J.-H., and No, W. G. (2012). The effect of first wave mandatory xbrl reporting across the financial information environment. *Journal of Information Systems*, 26(1):127–153.

Liu, C., Wang, T., and Yao, L. J. (2014). Xbrl's impact on analyst forecast behavior: An empirical study. *Journal of accounting and public policy*, 33(1):69–82.

Peri, G., Shih, K., and Sparber, C. (2015). Foreign and native skilled workers: What can we learn from h-1b lotteries? Technical report, National Bureau of Economic Research.

Sharifkhani, A. (2018). Immigration policy and equity returns: Evidence from the h-1b visa program. In *31st Australasian Finance and Banking Conference*.

van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2020). Man vs. machine learning: The term structure of earnings expectations and conditional biases. Technical report, National Bureau of Economic Research.

Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *The Review of Economic Studies*, 77(2):779–805.

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

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica: Journal of the Econometric Society*, pages 1415–1430.

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.

Appendix

The results are proved in a multi-asset setting, in which the assets' payoffs are independent and normally distributed. The capital Greek letters in this section (multi-asset) correspond to the parameters denoted in little Greek letters in section 2 (one risky asset).

Proof of Poposition 1

Proof. (a)Derivation of excess return

The derivation of $E[R]$ follows Kacperczyk et al. (2016) closely.

Conjecture that that the prices provide an unbiased signal about f, $\eta_p = f + \epsilon_p$, where $\epsilon_p \sim N(0, \Sigma_p)$. In a linear equilibrium where p is a linear function of investors' signals and asset supply,

$$p = \frac{1}{r}(A + Bf + Cx)$$

with A, B and C to be determined.

Based on private signals and prices, investors update their belief,

$$\hat{\mu}_i = E[f|\eta_{si}, p] = \hat{\Sigma}_i(\Sigma^{-1}\mu + \Sigma_p^{-1}\eta_p + K_i\eta_{si})$$

$$\hat{\Sigma}_i = var[f|\eta_{si}, p] = (\Sigma^{-1} + \Sigma_p^{-1} + K_i)^{-1}$$

First order conditions then give their holding conditional on signal realizations:

$$q_i = \frac{1}{\rho}\hat{\Sigma}_i^{-1}(\hat{\mu}_i - pr)$$

Market clearing gives

$$\begin{aligned}A =&\bar{\Sigma}(\Sigma^{-1}\mu - \rho\bar{x})\\ B =&I - \bar{\Sigma}\Sigma^{-1}\\ C =&- \rho\bar{\Sigma}(I + \frac{1}{\sigma_x^2\rho^2}\bar{K})\\ \Sigma_p^{-1} =&(\sigma_x^2 B^{-1}CC'B^{-1'})^{-1} = \frac{1}{\sigma_x^2\rho^2}\bar{K}\bar{K}^T,\end{aligned}$$

$\bar{K} = \int K_i di$ and $\bar{\Sigma}^{-1} = \int \hat{\Sigma}_i^{-1} di = \Sigma^{-1} + \Sigma_p^{-1} + \bar{K}$.

Using the expression of p, we can write $f - pr$ as

$$\begin{aligned}f - pr =&(I - B)f - Cx - A\\ =&\bar{\Sigma}[\Sigma^{-1}z + \rho(I + \frac{1}{\sigma_x^2\rho^2}\bar{K}^{-1})x] + \rho\bar{\Sigma}\bar{x}\\ =&V^{\frac{1}{2}}u + w\end{aligned}$$

where $V = \bar{\Sigma}[\rho^2\sigma_x^2 I + \bar{K} + \bar{\Sigma}^{-1}]\bar{\Sigma}$, $u \sim N(0,1)$, and $w = \rho\bar{\Sigma}\bar{x}$.

The market average holding is then given by

$$\begin{aligned}\bar{q} =&\frac{1}{\rho}\int \hat{\Sigma}_i^{-1}(\hat{\mu}_i - pr)di\\ =&\frac{1}{\rho}[\bar{K}z + \Sigma_p(z + \epsilon_p) + \bar{\Sigma}^{-1}(\mu - pr)]\end{aligned}$$

Investor i's excess holding

$$\begin{aligned}q_i - \bar{q} =&\frac{1}{\rho}[K_i\epsilon_{si} + (\hat{\Sigma}_i^{-1} - \bar{\Sigma}^{-1})(u + z - pr)]\\ =&\frac{1}{\rho}[K_i\epsilon_{si} + \Delta_i(V^{\frac{1}{2}}u + w)]\end{aligned}$$

where $\Delta_i = \bar{\Sigma}_i^{-1} - \bar{\Sigma}^{-1} = K_i - \bar{K}$. The last equality follows from $u + z - pr = f - pr$ and the expression of $f - pr$.

It is straightforward to check that all the matrices above are diagonal. With the expression of $q_i - \bar{q}$ and $f - pr$, investor's excess return can be easily calculated.

$$E_i[R] = E[(q_i - \bar{q})^T(f - pr)] = \rho(\bar{x}^T\bar{\Sigma}\Delta_i\bar{\Sigma}\bar{x}) + \frac{1}{\rho}Tr(\Delta_i V)$$

where $Tr(\cdot)$ is the trace operator.

Since all matrices are diagonal, the excess return of asset j for an investor with (n_p, n_f) is given by

$$E[R_j(n_p, n_f)] = \rho\bar{x}_j^2\bar{\sigma}_j^2(K(D(c_j, n_p), n_f) - \bar{K}_j) + \frac{1}{\rho}(K(D(c_j, n_p), n_f) - \bar{K}_j)v_j$$

where $\bar{x}_j$ is the jth element of $\bar{x}$. $\bar{\sigma}_j = (\bar{\Sigma})_{jj}$ and $v_j = (V)_{jj}$ are the jth diagonal element of $\bar{\Sigma}$ and V respectively.

The difference of the excess return on asset j relative to the base type investors is

$$\begin{aligned}\Delta R_j(n_p, n_f) =& E[R_j(n_p, n_f)] - E[R_j(\underline{n}_p, \underline{n}_f)] \\ =& \rho\bar{x}_j^2\bar{\sigma}_j^2(K(D(c_j, n_p), n_f) - K_j(\underline{n}_p, \underline{n}_f)) + \frac{1}{\rho}(K(D(c_j, n_p), n_f) - K_j(\underline{n}_p, \underline{n}_f))v_j\end{aligned}$$

The change of the difference after the shock that decreases c_j is

$$\gamma_j(n_p, n_f) = \Delta R'_j(n_p, n_f) - \Delta R_j(n_p, n_f)$$

If the shock is only on asset j, then $K_{-j}(D(c_{-j}, n_p), n_f)$ is not affected. So are $\bar{\sigma}_{-j}$ and v_{-j}. Thus ΔR_{-j} does not change and $\gamma_{-j}(n_p, n_f) = 0$.

For asset j, a decrease in c_j implies an increases in D, which in turn implies that $\bar{K}_j$ increases. So does the precision of price signal σ_{pj}^{-1}. Since $\bar{\sigma}_j^{-1} = \sigma_j^{-1} + \bar{K}_1 + \sigma_{pj}^{-1}$, $\bar{\sigma}_1$ decreases.

It remains to check how v_j changes. We only need to consider

$$\begin{aligned}
\frac{\partial v_j}{\partial \bar{K}_j} =& \frac{\partial}{\partial \bar{K}}\left(\frac{1}{\sigma^{-1}+\sigma_p^{-1}+\bar{K}}+\frac{\rho^2\sigma_x^2+\bar{K}}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2}\right)\\
=& -\frac{1}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2}+\frac{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2-2(\rho^2\sigma_x^2+\bar{K})(\sigma^{-1}+\sigma_p^{-1}+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^4}\\
=& -\frac{2(\rho^2\sigma_x^2+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^3}\\
&<0
\end{aligned}$$

The subscript j is omitted on the RHS.

Consider

$$\begin{aligned}
\frac{\partial(K(D(c_j,\underline{n}_p),\overline{n}_f)-K(D(c_j,\underline{n}_p),\underline{n}_f))}{\partial c_j} =& \frac{\partial}{\partial c_j}\int_{\underline{n}_f}^{\overline{n}_f}\frac{\partial K(D(c_j,\underline{n}_p),n_f)}{\partial n_f}dn_f\\
=& \int_{\underline{n}_f}^{\overline{n}_f}\frac{\partial^2 K(D(c_j,\underline{n}_p),n_f)}{\partial n_f \partial n_p}\cdot\frac{\partial D(c_j,\underline{n}_p)}{\partial c_j}dn_f
\end{aligned}$$

If $\frac{\partial^2 K(D,n_f)}{\partial n_f \partial D} \leq 0$, i.e., D and n_f are substitutes, then $\frac{\partial(K(D(c_j,\underline{n}_p),\overline{n}_f)-K(D(c_j,\underline{n}_p),\underline{n}_f))}{\partial c_j} \geq 0$. Combining with the results that $\bar{\sigma}_1$ and v_1 decrease, we conclude that $\gamma_j(\underline{n}_p,\overline{n}_f) < 0$ when c_j decreases.

If $\frac{\partial^2 K(n_p,n_f)}{\partial n_f \partial n_p} > 0$, i.e., n_p and n_f are complements, then $\frac{\partial(K(D(c_j,\underline{n}_p),\overline{n}_f)-K(D(c_j,\underline{n}_p),\underline{n}_f))}{\partial c_j} < 0$. In this case, the sign of $\gamma_1(\underline{n}_p,\overline{n}_f)$ can not be pinned down in general. However, it is clear that $\frac{\partial^2 K(n_p,n_f)}{\partial n_f \partial n_p} > 0$ is a necessary condition for $\gamma_j(\underline{n}_p,\overline{n}_f) \geq 0$ if c_j decreases.

□

Proof of Proposition 2

Proof. It follows directly from the last part of the previous proof, which shows that if $\hat{\sigma}(\underline{n}_p,\overline{n}_f)$ decreases more than $\hat{\sigma}(\underline{n}_p,\underline{n}_f)$, then it must be that $\frac{\partial^2 K(n_p,n_f)}{\partial n_f \partial n_p} > 0$. If $\frac{\partial^2 K(n_p,n_f)}{\partial n_f \partial n_p} \leq 0$, then the opposite is true. □

Proof of Proposition 3

Proof. To calculate the standard deviation of excess holding on asset j by investors of type (n_p, n_f), it is useful to consider the following holding dispersion. For a given investor i,

$$\begin{aligned} E[(q_{ij}-\bar{q}_j)^2] =& \frac{1}{\rho^2}E[(\Delta_{ij}(v_j^{\frac{1}{2}}u_j+w_j)+K_{ij}\epsilon_{ij})^2] \\ =& \frac{1}{\rho^2}E[(K_{ij}-\bar{K}_j)^2(v_j^{\frac{1}{2}}u_j+w_j)^2+2(K_{ij}-\bar{K}_j)(v_j^{\frac{1}{2}}u_j+w_j)K_{ij}\epsilon_{ij}+K_{ij}^2\epsilon_{ij}^2] \\ =& \frac{1}{\rho^2}[(v_j+\rho^2\bar{\sigma}_j^2\bar{x}_j^2)(K_{ij}-\bar{K}_j)^2+K_{ij}] \end{aligned}$$

The first equality follows from the expression of $q_i - \bar{q}$. $\Delta_{ij} = K_{ij} - \bar{K}_j$ has been used in the second line. The third equality follows from the fact that $u \sim N(0,1)$, $w_j = \rho\bar{\sigma}_j\bar{x}_j$ and $E[\epsilon_{ij}] = K_{ij}^{-1}$. USing these results,

$$E[(q_{ij}-\bar{q}_j)^2] - E[(q_{ij}-\bar{q}_j)]^2 = \frac{1}{\rho^2}[v_j(K_{ij}-\bar{K}_j)^2+K_{ij}]$$

Since investors belonging to the same type have the same K_{ij}, we have (3)

$$\delta_j(n_p,n_f) = \frac{1}{\rho}\sqrt{v_j(K_j(D(c_j,n_p),n_f)-\bar{K}_j)^2+K_j(n_p,n_f)}$$

Noting that $K_j(n_p,n_f)$ increases while v_j decreases after the shock. We conclude that if δ_j decreases then the first term under the square root dominates. In addition, as $\frac{\partial v_j}{\partial \bar{K}_j} = -\frac{2(\rho^2\sigma_x^2+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^3}$, the magnitude of the change due to an infinitesimal increase in K becomes unbounded when σ_x goes to infinity. Thus when σ_x is large enough, we only need to consider the first term,

$$\delta_j(n_p,n_f) \approx \frac{1}{\rho}\sqrt{v_j}|K_j(n_p,n_f)-\bar{K}_j| \tag{6}$$

For the base type investors, we have immediately that

$$\delta_j(\underline{n}_p, \underline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\underline{n}_p, \underline{n}_f)),$$

since they have the smallest K. $\delta_j(\underline{n}_p, \underline{n}_f)$ is likely to decrease after the shock, due to a smaller v_j and a possibly smaller distance to the average precision, $\bar{K}_j - K(\underline{n}_p, \underline{n}_f)$. The latter comes from the concavity of the production function.

As the dispersion for the untreated asset is not affected, we only need to consider the changes among different type of investors for the treated asset. If $K(\overline{n}_p, \overline{n}_f)$ is large enough such that $K(\overline{n}_p, \underline{n}_f) < \bar{K}_j$ and $K(\underline{n}_p, \overline{n}_f) < K_j$, (6) becomes

$$\delta_j(\overline{n}_p, \underline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\overline{n}_p, \underline{n}_f))$$

and,

$$\delta_j(\underline{n}_p, \overline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\underline{n}_p, \overline{n}_f))$$

The impact of the shock on the difference in dispersion the difference between the finance-intensive type and the base type is given by

$$\begin{aligned}\frac{\partial(\delta_j(\underline{n}_p, \overline{n}_f) - \delta_j(\underline{n}_p, \underline{n}_f))}{\partial c_j} =& \frac{1}{2\rho\sqrt{v_j}}\frac{\partial v_j}{\partial c_j}(K(\underline{n}_p, \underline{n}_f) - K(\underline{n}_p, \overline{n}_f)) \\ &+ \frac{1}{\rho}\sqrt{v_j}\frac{\partial(K(\underline{n}_p, \underline{n}_f) - K(\underline{n}_p, \overline{n}_f))}{\partial c_j}\end{aligned}$$

The first term is negative. The second term is negative in case of substitution and positive in case of complementarity. Therefore, a negative change in the difference following a decrease in c_j implies complementarity.

□

Table 1: Summary statistics. All continuous variables are winsorized at the top and bottom one percent to mitigate the influence of extreme values

Variables	N	mean	std	p25	median	p75
JobPerInst(IT)	421	10.24	115.51	0.00	0.00	1.00
JobPerInst(finance)	421	8.01	74.91	0.00	1.00	2.00
Assets(base)	227	1981.37	4395.51	130.09	549.92	1800.86
Assets(IT)	79	11920.84	33407.91	526.73	1992.71	8372.48
Assets(finance)	133	447.86	2505.98	29.12	81.70	203.06
Assets(Bi)	38	1381.04	4168.60	29.23	109.77	544.57
MarketCap	496873	5.63	19.83	0.38	1.09	3.10
RetHolding	486935	-1.29	11.02	-2.53	-0.16	0.62
RetTrading	357328	0.50	16.17	-0.54	0.00	0.67
Fraction	96626	0.03	0.05	0.00	0.01	0.05
Fraction_Scaled	96627	1.50	5.46	0.06	0.24	0.80
InstOwn	96609	0.59	0.28	0.36	0.64	0.83
Book-to-market	95146	0.82	0.73	0.36	0.65	1.05
Debt-to-equity	95182	2.29	3.96	0.41	1.02	2.44
Dispersion	74319	0.40	1.84	0.01	0.04	0.19
AveMarketCap	3289	0.34	0.85	0.01	0.05	0.25
NumOfStocks	3122	131.20	259.62	9.00	30.00	120.00
Turnover	2837	0.16	0.12	0.06	0.12	0.23

Table 2: Classification for the twenty largest institutions.

IT-intensive	Finance-intensive	Bi-intensive
LORD, ABBETT & CO. LLC	BARCLAYS BANK PLC	MERRILL LYNCH & CO INC
LAZARD CAPITAL MARKETS LLC	MOORE CAPITAL MANAGEMENT, INC.	UBS SECURITIES LLC
KEYBANK NATIONAL ASSOCIATION	BALYASNY ASSET MANAGEMENT LP	SG AMERICAS SECURITIES, LLC
D. E. SHAW & CO., L.P.	ROYAL BANK OF CANADA	BAIN CAPITAL, LLC
CITIGROUP INC	FIRST QUADRANT L.P.	BLACKSTONE GROUP
GENERAL ELECTRIC COMPANY	MASON CAPITAL MANAGEMENT	SOROS FUND MANAGEMENT, L.L.C.
RCM CAPITAL MANAGEMENT LLC	VISIUM ASSET MANAGEMENT, LP	BRIDGEWATER ASSOCIATES INC.
RENAISSANCE TECHNOLOGIES CORP.	TUDOR INVESTMENT CORPORATION	TPG CAPITAL, L.P.
LOEWS CORPORATION	ROCHDALE INVESTMENT MGMT LLC	WOLVERINE ASSET MGMT, L.L.C.
CREDIT SUISSE SECS (USA) LLC	COBALT CAPITAL MGMT, INC.	DAVIDSON KEMPNER CAP MGMT L.L.
AQR CAPITAL MANAGEMENT, LLC	TFS CAPITAL LLC	ENVESTNET ASSET MGMT, INC.
FISHER INVESTMENTS	TREMBLANT CAPITAL GROUP	INTEL CORPORATION
COHEN & STEERS CAP MGMT, INC.	THIRD POINT LLC	QVT FINANCIAL LP
JACOBS LEVY EQUITY MGMT, INC.	SIRIOS CAPITAL MGMT, L.P.	ATTICUS CAPITAL, L.P.
HIGHBRIDGE CAPITAL MGMT, LLC	U.S. GLOBAL INVESTORS, INC.	CTC LLC
TWO SIGMA INVESTMENTS, LLC	SANDLER CAPITAL MANAGEMENT	NORTHCOAST ASSET MGMT LLC
ACADIAN ASSET MANAGEMENT LLC	CAPSTONE INVT ADVISORS, LLC	NUVEEN LLC
GAMCO INVESTORS, INC.	RHO CAPITAL PARTNERS, INC.	MARINER INVESTMENT GROUP LLC
BRANDES INVT PARTNERS, LP	BRAHMAN CAPITAL CORP.	TRAXIS PTNR LLC
ARROWSTREET CAPITAL, L.P.	CYRUS CAPITAL PARTNERS, L.P.	ELLINGTON MGMT GROUP, L.L.C.

Table 3: Results of regression on institutional characteristics. $y_{it} = \alpha + \sum_{\theta} \beta_\theta Type_{i\theta t} + \alpha_t + Controls_{it}$. In column (1) the dependent variable is the institution's asset turnover. The dependent variable in column (2) is the log value of the average market capitalization of portfolio stocks of an institution at time t. The average is weighted by portfolio weights. Column (3) uses the log value of the number of portfolio stocks as the dependent variable. In all the regressions, I control for institution size. Standard errors are clustered at quarter and institution level.

	(1) Turnover	(2) AveMarketCap	(3) LogNumStocks
Bi-intensive	0.00972	0.64836	-0.24438
	(0.320)	(1.225)	(-0.421)
IT-intensive	0.13873***	-0.58262	0.70796
	(3.136)	(-1.034)	(1.210)
Finance-intensive	-0.08354*	-0.95361*	-0.41504
	(-2.117)	(-1.910)	(-0.982)
Control	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	2,837	3,289	3,122
R^2	0.06	0.62	0.40

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Results of regression on excess returns(bps). $r_{ijqc} = \sum_{\theta} \gamma_\theta \, Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^r \, XBRL_{jc} \times Post_q +$ Institution FE $+$ Stock-Type-Cohort FE $+$ Type-Period-Cohort FE $+$ $Controls_{ijqc}$. Period q is measured as the number of quarters relative to the implementation XBRL mandate in each cohort. The dependent variable in the first two columns is excess holding return. The dependent variable in columns (3) and (4) is the excess trading return. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$. Controls include asset under management, institution turnover, and stock market capitalization. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)	(4)
	Holding	Holding	Trading	Trading
XBRL × Post	-0.67671		-0.04242	
	(-0.565)		(-0.167)	
XBRL × Post × IT	-0.60411***	-0.76533**	-0.41631**	-0.73453***
	(-3.679)	(-3.098)	(-2.715)	(-4.484)
XBRL × Post × Finance	0.82200**	1.10378*	2.19206***	1.84578**
	(2.574)	(1.996)	(3.566)	(2.729)
XBRL × Post × Bi-intensive	-0.50870*	-0.69297*	-0.04490	-0.31821
	(-2.053)	(-1.847)	(-0.104)	(-0.811)
Controls	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes	Yes
Observations	470,871	468,850	346,288	343,846
R^2	0.24	0.51	0.36	0.73

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Results of regression on stock ownership. $Ownership_{j\theta qc} = \sum_{\theta}\gamma^s_\theta \; Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^s \; XBRL_{jc} \times Post_q$+Stock-Type-Cohort FE+Type-Period-Cohort FE+$Controls_{ijqc}$. The dependent variable in the first two columns is the fraction of total share outstanding owned by each type of investor. In column (3) and column (4), the fraction is divided by the total assets of the type. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$ and Stock FE. Control variables include stock market capitalization, book-to-market ratio, debt-to-equity ratio, total institutional ownership and average asset under management of each type. Standard errors are clustered at stock and quarter-cohort level.

	(1)	(2)	(3)	(4)
XBRL × Post	0.00125		0.22367**	
	(1.518)		(2.882)	
XBRL × Post × IT	-0.00657***	-0.00630***	-0.22843**	-0.20367**
	(-6.207)	(-5.900)	(-3.102)	(-2.612)
XBRL × Post × Finance	-0.00108	-0.00134	0.88590***	0.86258**
	(-1.267)	(-1.524)	(3.180)	(2.975)
XBRL × Post × Bi-intensive	-0.00173*	-0.00154	-0.32760***	-0.28570***
	(-1.891)	(-1.621)	(-3.690)	(-3.208)
Controls	Yes	Yes	Yes	Yes
Stock FE	Yes	No	Yes	No
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes	Yes
Observations	87,642	84,972	87,662	84,817
R^2	0.95	0.97	0.77	0.84

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Results of regression on stock holding dispersion, $\delta_{j\theta qc} = \sum_{\theta}\gamma_{\theta}\ Type_{i\theta c} \times XBRL_{jc} \times Post_{qc} + \beta^{\delta}\ XBRL_{jc} \times Post_{qc}$+Stock-Type-Cohort FE+Type-Period-Cohort FE+ $Controls_{ijqc}$. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$ and Stock FE. Control variables include stock market capitalization, book-to-market ratio, debt-to-equity ratio, total institutional ownership, and average asset under management of each type. In columns (3) and (4), Stock-Type-Cohort FE is replaced by Treated-Type-Cohort FE. Standard errors are clustered at stock and quarter-cohort level.

	(1)	(2)	(3)	(4)
XBRL × Post	-0.09160		-0.09992*	
	(-1.424)		(-2.013)	
XBRL × Post × IT	0.09408	0.09853	0.10746*	0.11536*
	(1.486)	(1.528)	(2.099)	(2.072)
XBRL × Post × Finance	-0.09361	-0.05572	-0.15142**	-0.14356**
	(-0.691)	(-0.409)	(-2.473)	(-2.249)
XBRL × Post × Bi-intensive	0.01195	-0.00466	0.12168*	0.13588**
	(0.117)	(-0.050)	(2.192)	(2.712)
Controls	Yes	Yes	Yes	Yes
Stock-Cohort FE	Yes	No	Yes	No
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	No	No
Treated-Type-Cohort FE	No	No	Yes	Yes
Observations	64,599	60,183	66,781	62,813
R^2	0.82	0.87	0.30	0.36

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Assets(millions) by investor type in the matched sample.

	Base_{it}	IT	Base_{fin}	Finance	Base_{bal}	Balanced
count	225.00	73.00	201.00	132.00	199.00	36.00
mean	2071.44	4911.52	858.13	446.33	1330.27	822.18
std	4510.77	8051.18	1027.20	2515.71	2812.11	2323.74
min	1.60	2.09	1.60	0.34	1.60	1.58
median	550.84	1513.23	395.50	75.85	463.63	115.86
max	46805.44	45549.32	6491.78	28686.12	28051.74	13495.52

Table 8: Regressions on excess holding returns (matched sample). $r_{ijtc} = \gamma_\theta \; Type_{i\theta c} \times XBRL_{jc} \times Post_t + XBRL_{jc} \times Post_t + Controls_{ijtc}$+Institution FE+Stock-Type-Cohort FE+ Type-Period-Cohort FE. Standard errors are clustered at institution and quarter-cohort level.

	(1) IT	(2) Finance	(3) Bi-intensive
IT × XBRL × Post	-0.81285** (-2.871)		
Finance × XBRL × Post		1.28249** (3.083)	
Bi × XBRL × Post			-0.18505 (-0.417)
Controls	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Stock-Period-Cohort FE	Yes	Yes	Yes
Type-Period-Cohort FE	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes
Observations	307,198	130,744	172,247
R^2	0.49	0.37	0.43

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regressions on excess trading returns (matched sample). $r_{ijqc} = \gamma_\theta \ Type_{i\theta c} \times XBRL_{jc} \times Post_{qc} + XBRL_{jc} \times Post_{qc} + Controls_{ijqc}$ + Institution FE + Stock-Type-Cohort FE + Type-Period-Cohort FE. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)
IT × XBRL × Post	-0.34153*** (-3.477)		
Finance × XBRL × Post		1.37825*** (4.643)	
Bi × XBRL × Post			-0.29868 (-1.322)
Controls	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Stock-Period-Cohort FE	Yes	Yes	Yes
Type-Period-Cohort FE	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes
Observations	226,084	87,109	113,663
R^2	0.54	0.43	0.47

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Results on portfolio returns(bps). $r_{ijqc} = \underset{\theta}{\Sigma}\gamma_\theta \; Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^r \; XBRL_{jc} \times Post_q +$ Institution FE + Group-Type-Cohort FE + Type-Period-Cohort FE + $Controls_{ijqc}$. Period q is measured as the number of quarters relative to the implementation XBRL mandate in each cohort. The dependent variable in the first two columns is excess holding return. The dependent variable in columns (3) and (4) is the excess trading return. In columns (2) and (4), I control for Group-Period-Cohort FE, which absorbs $XBRL \times post$. Controls include asset under management and institution turnover. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)	(4)
	Holding	Holding	Trading	Trading
XBRL × Post	0.38		1.27	
	(0.44)		(1.75)	
XBRL × Post × IT	-0.75	-0.42	-2.22**	-1.89**
	(-0.70)	(-0.39)	(-3.08)	(-2.74)
XBRL × Post × Finance	1.64	1.86	1.18	1.34
	(1.22)	(1.46)	(1.08)	(1.21)
XBRL × Post × Bi-intensive	0.13	0.15	-0.53	-0.61
	(0.08)	(0.10)	(-0.21)	(-0.24)
Controls	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes
Group-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
group-Type-Cohort FE	Yes	Yes	Yes	Yes
Observations	4,893	4,893	4,064	4,064
R^2	0.29	0.29	0.34	0.34

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

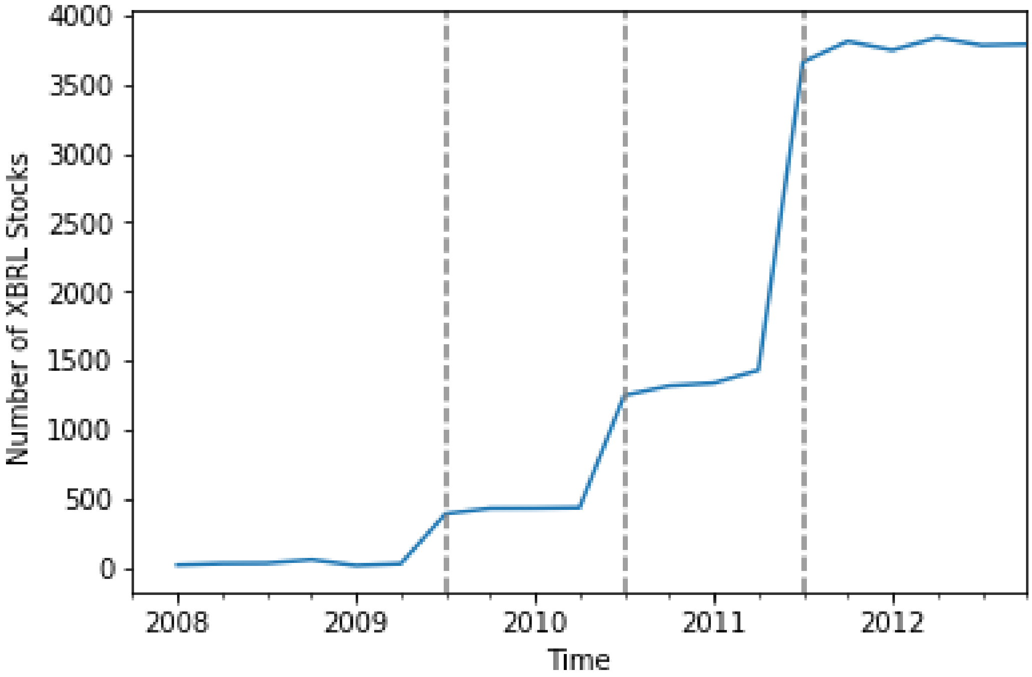

Figure 2: Number of stocks that complied with the XBRL mandate in each quarter.

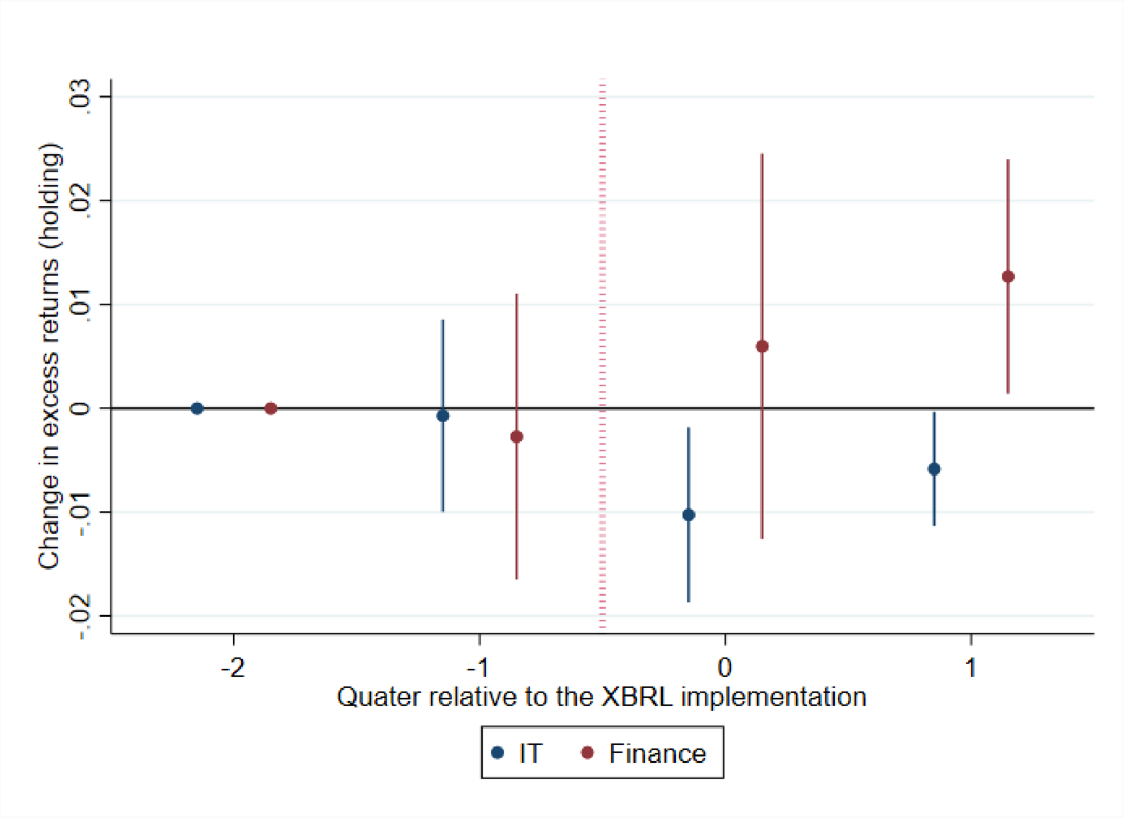

Figure 3: Plot of differential impact for holding returns (in percentage points). It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.

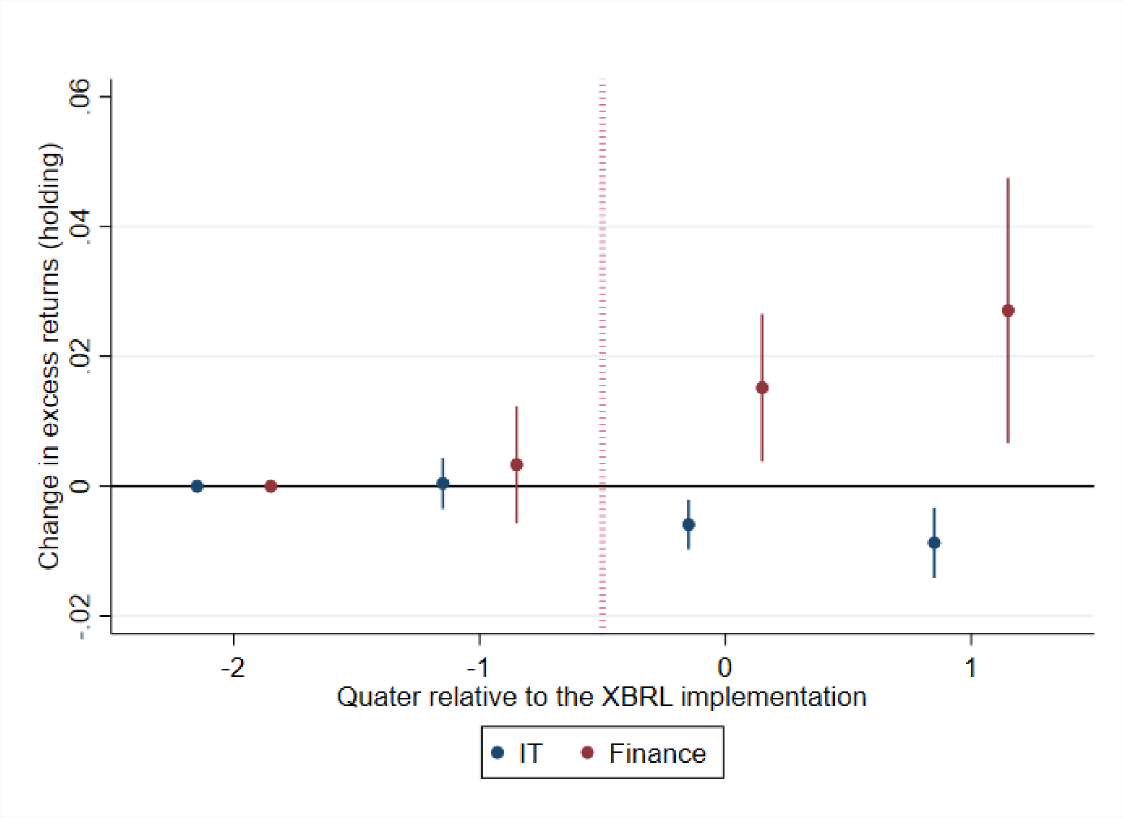

Figure 4: Plot of differential impact for trading returns (in percentage points). It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.

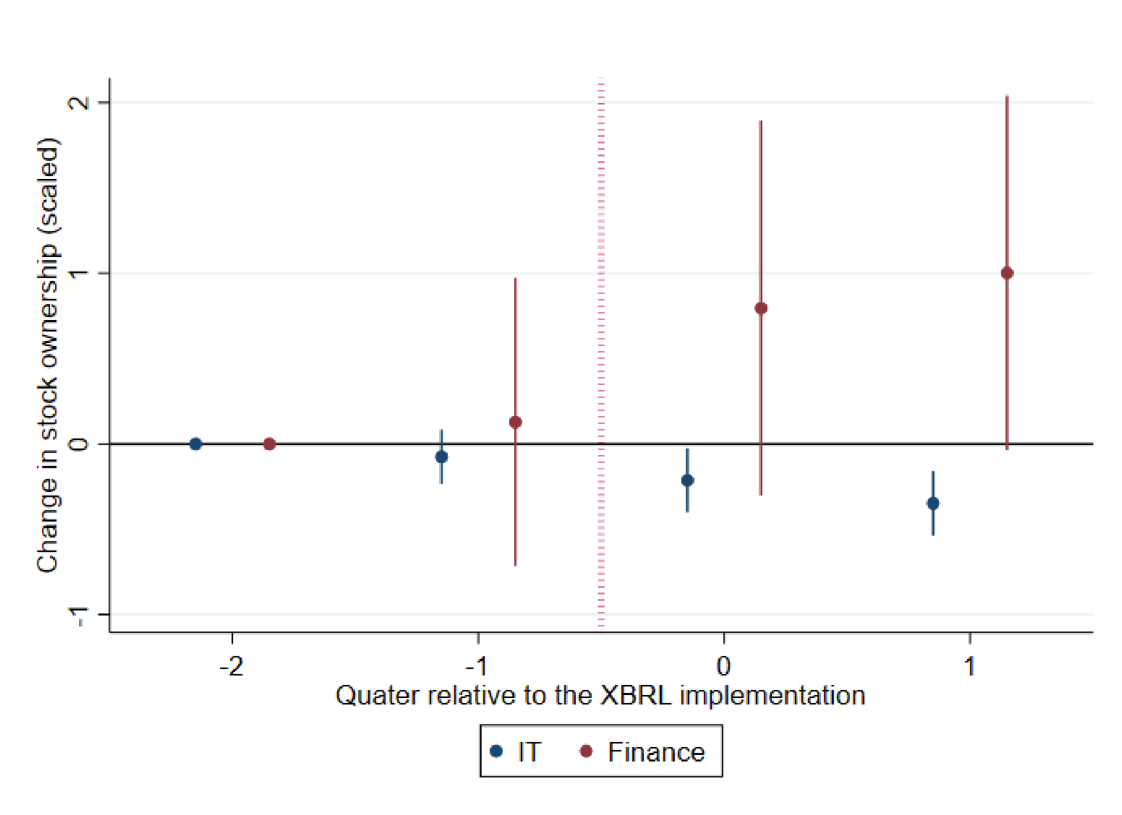

Figure 5: Plot of differential impact for stock ownership. It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.