Using Digitalized Financial Statements to Identify Firms Innovating without R&D

Hyuna Park*

January 8, 2023

Abstract

Innovation drives growth and creates value, but measuring it is challenging because many intangible investments are expensed rather than capitalized under the current accounting rules. Prior research uses R&D expenditures due to the limited data availability, but innovative investments should be more broadly defined than R&D. I develop new methods to identify innovative firms using i) XBRL tags that work like barcodes in digitalized financial statements and ii) machine learning tools applied to financial statement texts. The tags for communications, information technology, and data processing indicate innovative investments in the finance industry, for example. I apply textual analysis to industries with neither R&D nor relevant XBRL tags, such as retail, and estimate that they have over $240 billion of unrecorded intangible assets. I show that the investment gap reported in prior research becomes insignificant when adjusting for the estimated intangible investments.

Keywords: Unrecorded Intangibles, Innovation, eXtensible Business Reporting Language (XBRL)
JEL Classification: O32, C88, G30, M49

* Hyuna Park, is Herb Kurz Chair in Finance and Risk Management at Brooklyn College of the City University of New York; e-mail hyuna.park38@brooklyn.cuny.edu, phone (413) 348-9116. I thank Aparna Gupta, Ariel Avshalom, Aniruddha Ganguly, Soon Chun and Games Jeller for helpful discussions and information on textual analysis and machine learning tools, Gabriel Rudy and Marine Kocollari for research assistance, and Soku Byoun, Sang Mook Lee, Steve Lim, SungIn Moon, and participants of the Korea-America Finance Association (KAFA) brown bag seminar, Koppelman School of Business research seminar, and the seminar of the alumni association in Minnesota of Seoul National University (SNUAA – MN) for helpful discussions on innovation and intangible investments.

1. Introduction

Innovative investments drive growth and value creation; thus, accurate measurement is critical for investors, policymakers, and other stakeholders. Are R&D expenditures in Compustat (XRD) used in most previous research the best measure of innovative investments? The answer is no because many firms without XRD show clear evidence of innovation. For example, over 60,000 US patents were granted from 1974 to 2020 to companies with no XRD. These patents had a total market value of over $2.3 trillion, deflated to 1982 dollars using the consumer price index (CPI), according to Kogan *et al.* (2017). They received over 700,000 citations, as shown in Table 1.

How can we measure innovative investments of firms that do not have XRD? The main objective of this paper is to answer this question, and I propose two methods that use digitalized financial statements as alternative data sources**. The first method uses eXtensible Business Reporting Language (XBRL) tags.** The tags work like barcodes in digitalized financial statements and have been freely available for download from the following US Securities and Exchange Commission (SEC) website since 2009: SEC.gov | Financial Statement Data Sets

For example, JPMorgan (Ticker: JPM) has over nine hundred US patents granted since 1983, and the aggregate value of those patents is over $250 billion, deflated to 1982 dollars. JPM does not have any R&D expenditures in Compustat, like many other financial institutions. Still, their XBRL tags include us-GAAP: CommunicationsAndInformation Technology, and the expenses were $9.94 billion in 2021, as shown in the following website:https://www.sec.gov/ix?doc=/Archives/edgar/data/19617/000001961722000272/jpm-20211231.htm .

As shown in Table 1, firms filing patents without XRD are from many industries, and the finance, foods, and telecommunications industries are most noteworthy. Among the $2.3 trillion market value of patents in Table 1, $975 billion (42%) belongs to financial institutions with 2-digit SIC codes 60 – 63. After reviewing all monetary XBRL tags of financial institutions, I found that these firms use Communication (COM), CommunicationsAndInformationTechnology (CIT), and InformationTechnologyAndDataProcessing (ITDP) tags to report expenses related to technology and data.

I regress patent market values on these expenses and find that the coefficients are highly significant with *t*-statistics 6.95 ~ 13.05. These costs are regarded as "short-term" operating expenses under the US Generally Accepted Accounting Principles (GAAP) and thus not recorded as assets on balance sheets[1], but the regressions using patent data confirm that these expenses generate "long-term" benefits. The results show that the first tool works well for firms that report innovative investments using separate XBRL tags, but we need a different method for those who do not have such tags.

That is why I propose **the second method that uses textual analysis** to identify firms with neither XRD nor relevant XBRL tags but explain their innovative investments in words in their financial statements. For example, many firms in the retail industry have invested significantly in innovation for digital transformation since the rapid growth of the Internet started in the mid-1990s. However, many retailers have neither XRD nor separate XBRL tags showing these investments. They report these innovative investments combined with other

[1] Kothari et al. (2002) point out that the high uncertainty about the future benefits of intangibles is the rationale behind the immediate expensing decision. Many challenges accountants face when they value intangible assets lead to the issues of “conservative accounting biases in book value,” “unrecorded intangible assets,” and “unverifiable fair value estimates.” For more details, see Beaver and Ryan (2005), Lev and Zarowin (1999), Lev and Radhakrishnan (2005), Penman and Zhang (2002), Park (2019, 2022), and Ramanna and Watts (2012).

expenses, such as utility bills as part of selling, general and administrative expenses (XSGA), or costs of goods sold (COGS).

For example, Starbucks (Ticker: SBUX) has 34 US patents granted since 1996, and the most valuable one is "*Beverage preparation systems and methods*" shown in https://patents.google.com/patent/US10531761B2/en?oq=10531761. The grant date of this patent is 01/14/2020, the expiration date is 05/20/2037, and $603 million is the estimated market value. Numbers in financial statements do not show any investments in the innovation because internally developed intangibles are not capitalized under the US GAAP, but financial statement texts explain them: *"Starbucks owns and has applied to register numerous trademarks and service marks …trademarks are generally valid and may be renewed indefinitely … We own numerous copyrights for items such as product packaging, promotional materials, in-store graphics and training materials…hold patents on certain products, systems and designs which have an average remaining useful life of approximately seven years"* as shown in https://www.sec.gov/ix?doc=/Archives/edgar/data/ 829224/000082922422000058/sbux-20221002.htm#ia75cc8f98747496589a 1ed7893374c6c_43

I use natural language processing (NLP) and machine learning tools such as Word2Vec (W2V) to extract information on innovative investments from financial statement texts. I develop an intangibles dictionary (iDic) by collecting seed words on intangibles from documents of accounting firms advising on intellectual property in mergers and acquisitions and sample 10Ks of firms investing in intangibles. I use W2V to expand the vocabulary for each industry around the seed words by iteration.

I choose the retail industry as an example because most retailers do not report any R&D expenditures, but they have invested significantly in innovation for digital transformation during the past three decades and explain it in their 10Ks. I apply iDic to retail 10Ks and find that retailers whose iDic score increased one percentage point from the previous year spent seven percent more operating expenses than other retailers after controlling for retail subsectors, and the difference is significant at the one percent level. I use the iDic regression coefficients to estimate unrecorded intangible assets of retailers and show that intangible adjustment significantly affects profitability and investment shortfalls reported in prior research.

There is a growing concern in the economics literature reporting that investments of US firms fell relative to fundamentals such as cash flows and the market valuation of assets (Hall, 2014; Gutierrez and Phillippon, 2017; Alexander and Eberly, 2018; Grouzet and Eberly, 2018). They define the investment gap using tangible assets, and I extend the research by adding the estimated intangible investments to the analysis. I find that unrecorded intangible investments are the main reason for the investment shortfall of retailers reported in prior research. The main contribution of this paper is to show two new methods of measuring unrecorded intangible investments. These methods are most valuable for industries that invest heavily in intangibles but do not report R&D expenditures.

The rest of this paper is organized as follows. Section 2 explains how to use XBRL tags to measure innovative investments of firms that do not have XRD. Then I show the evidence of XBRL tags recorded as short-term operating expenses under the US GAAP generating long-term benefits using regressions of patent values on those expenses for prior years. Section 3 presents NLP and machine learning tools applied to financial statement texts to measure innovative investments of firms with neither XRD nor relevant XBRL tags. I explain methods to develop lists of words and phrases describing innovative investments and how to use them to measure the investment proportion of operating expenses in income statements. Section 4 presents cases where both XBRL tags and textual analysis are used to estimate intangible investments, and Section 5 concludes.

2. XBRL tags for firms innovating without XRD

2.1. XRD in Compustat and firms with patents and no XRD

I use the Center for Research in Security Prices (CRSP)-Compustat merged datasets released in July 2022 as the source of Compustat data. Prior research uses XRD in Compustat to measure investments in innovation, but over 50% of firms do not have R&D expenditures, and the proportion varies significantly across industries, as shown in Figure 1. For example, firms in health care and business equipment industries spend 10-14% of sales revenue on R&D, while most firms in finance do not have any R&D expenditures.[2]

However, no R&D expenditures do not mean no investments in innovation. As the patent data shows in Figure 2, finance firms have demonstrated significant innovation since the mid-1990s. It is when the digital economy started with the rapid growth of the Internet, according to the US Bureau of Economic Analysis (BEA) definition.[3] The proportions of finance patents have grown from near zero to 6 percent in numbers, 8 percent in market values, and 15 percent in citations during the past two decades.

The patent data in Figure 2 are from the GitHub data repository for Kogan *et al.* (2017). I thank them for making the data available for download from the following website: https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-

[2] Compustat reports multiple versions of data for the same year to standardize data items across companies and industries and CRSP classifies them using keysets. For example, keyset=1 is for standardized data and keyset=2 for originally reported values. The proportion of firms missing XRD is much higher when using the original data in keyset=2 than in the standardized values in keyset=1. Figure 1 presents the standardized data in keyset=1 and shows over 50 percent of missing XRD. The proportion would be even higher if keyset=2 were used.

[3] Digital Economy | U.S. Bureau of Economic Analysis (BEA) provides the definition and related research.

Extended-Data. They estimate the market value of each patent using the stock price movement on the grant days using Google Patents as the data source. They develop an automation script for downloading the patent data and a name-matching algorithm to check if the assignee of each patent is in the CRSP database for stock returns. They examine the trading volume data around the patent issue dates, find significant increases in turnover on the first two days following the announcement, and thus use the three-day window for estimation. When estimating the patent market value, they use idiosyncratic return, defined as each firm's return minus the market portfolio return, to sort out market movements from the impact of patents.

2.2. XBRL tags

Patent data shows firms investing in innovation without XRD in Compustat; thus, I use XBRL tags of these firms as an alternative data source for investments in innovation. The tags make it possible to read and process big data fast and accurately, just like using barcodes to keep track of information electronically. The SEC initially launched the XBRL reporting as a voluntary financial reporting program (VFP) in 2003 (Release No 33-8496). Then it became a requirement in 2009 (Release No 33-9002): Final Rule: Interactive Data to Improve Financial Reporting (sec.gov). In 2018, the SEC adopted amendments requiring Inline XBRL, meaning that companies prepare a single document that is both human-readable and machine-readable. For a sample financial statement that is not human-readable but for machine-reading, see the following: https://www.sec.gov/Archives/edgar/data/19617/0000019617180000 57/0000019617-18-000057.txt

We can download XBRL data from the following SEC website since 2009, updated quarterly: SEC.gov | Financial Statement Data Sets. Each quarterly data folder includes four text files, SUB, NUM, TAG, PRE, and a data manual, *The Financial Statements Data (PDF, 175 kb),* that explains the scope, organization, file formats, and definitions.

The submission data set (SUB) includes one record for each XBRL submission during the quarter and shows information about the submission and the filing entity, such as ADSH and CIK. ADSH is an accession number the SEC assigns to each submission to its Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. Central Index Key (CIK) is a ten-digit number the SEC assigns to each registrant that submits filings.

The number data set (NUM) includes one row for each amount for all line item values from each submission in SUB. As ADSH is in both datasets, I use it to link the two datasets. The tag data set (TAG) shows each tag's definitions, descriptions, versions, and other information. For example, a *CommunicationsAndInformationTechnology* tag is us-gaap/2020 version, the custom variable sets to zero because this tag is not for a specific company but for us-gaap, the datatype is monetary, and the definition is the amount of expense in the period for communications and data processing expense. I use the tag, custom, and datatype in this data set to sort out all monetary tags based on us-gaap.

The presentation data set (PRE) shows where each tag was presented in the primary financial statements. For example, the data set shows that stmt is IS for *CommunicationsAnd InformationTechnology* tags meaning that these tags appear in income statements. I use the ADSH, stmt, and tag columns in this dataset to sort out all tags in income statements.

To my knowledge, this paper is the first to use XBRL to analyze investments in innovation. XBRL research has a short history as the data became available recently, and Hoitash *et al.* (2021) review the literature. They point out the pros and cons of Compustat and XBRL data, such as Compustat distributing less granular and standardized data that may be different from what was originally reported in XBRL tags. Chychyla and Kogan (2015) analyze the discrepancies between Compustat data and XBRL records and find that 17 out of 30 analyzed variables in Compustat are significantly different from corresponding XBRL tag values. Dong *et al.* (2016) use the adoption of XBRL as a natural experiment to test the theoretical reasoning of underinvestment in the production of expensive firm-specific information. They find that the effect of XBRL adoption on stock return synchronicity is significant for complex firms that have financial information that is inherently more difficult to process.

2.3.XBRL tags related to innovation in the finance industry

I use the finance industry to show how to use XBRL tags to measure innovative investments because the industry has rapidly growing patents without XRD. I examined the monetary XBRL tags in the income statements of financial firms that have multiple patents and identified tags related to innovation. I used the SIC codes in the SUB dataset, *sic*, to identify financial firms and the *qtrs* variable in the NUM dataset to sort out annual data as it shows the count of the number of quarters represented by the tag data value.

There are 392,607 monetary XBRL tags in the income statements of financial firms (SIC 6000 – 6999) as of 2021 QTR 4 after deleting duplicates. Note that most firms include the previous two years of historical values with the current data for comparison, leading to the same tag values appearing multiple times in the XBRL datasets.

By examing the monetary income statement XBRL tags, I find that financial firms use Communication (COM), Communications AndInformationTechnology (CIT), and InformationTechnologyAndDataProcessing (ITDP) tags to report expenses related to technology and data. As shown in Table 2, COM has 769 firm-year observations, the average is $31.6 million, and the median is $1.8 million. The highest COM was Bank of America's $1.7 billion in 2012. CIT has 1,026 firm-year, the average is $324 million, and the median is $17.7 million. The highest CIT was JP Morgan's $10.3 billion in 2020. ITDP has 2,903 firm years, the average is $31.3 million, and the median is $2.6 million. The highest ITDP was Bank of America's $3.2 billion in 2018. Some companies report combinations of COM, CIT, and ITDP instead of one of them in the same year, and thus I also define the sum of the three as COMITDP, and show the summary statistics in Table 2.

I identified the three XBRL tags by examining the income statements of financial institutions with multiple patents, but we need formal tests of the relationship between these tags and patents. Thus I regress patent values on those tags and explain them in the following subsection.

2.4. Regressions of patent values on XBRL tags related to innovation

Table 3 presents regressions to test whether the prior expenditures of financial institutions on information technology and data processing explain the value of patents granted to them in the future. I match the patent data with XBRL tags using permno in CRSP, CIK in XBRL, and company description in the CRSP-Compustat merged database. The dependent variable is the total market value of patents filed for each firm-year. For each XBRL tag, COM, CIT, and ITDP, I calculate the total for three years prior to the patent filing year and use it as an explanatory variable.

As shown in Table 3, COM, CIT, ITDP all show highly significant results, with the *t*-statistics ranging from 6.95 to 10.06. As some firms report more than one of the three tags for the same year, I define COMITDP as the sum of COM, CIT, and ITDP and test it and find a stronger result (*t*-statistic for COMIDP = 13.05). These results confirm the "long-term" benefits of COM, CIT, and ITDP, even though they are currently regarded as "short-term" operating expenses and thus not allowed to be capitalized under the current accounting rules.

As COM, CIT, and ITDP data are from XBRL, and thus they are available only after 2009, we need another data source to test the relationship for earlier years. I found that XCOM in the *incomestatementfinancialannual* file of the CRSP-Compustat merged database is comparable to the sum of COM, CIT, and ITDP in XBRL by comparing these tag values with variables in Compustat. Among the 4,514 firm-years with COMITDP in the XBRL data, 2,961 has corresponding XCOM in Compustat, and the Pearson correlation coefficient between COMITDP and XCOM for the 2,961 firm-years is 0.9698.

Figure 3 shows that the proportion of financial firms reporting XCOM increased sharply from less than 5% to over 25% when the digital economy started in the mid-1990s. I regress the patent value on the previous three years of XCOM and find that the coefficient is highly significant (*t*-statistic = 11.15), as shown in Table 3. Overall, these results confirm that operating expenses for information technology of financial institutions generate long-term benefits in the form of patent market values even though balance sheets do not show these internally generated intangible assets.

When firms have separate XBRL tags for innovative investments such as COM, CIT, and ITDP in the finance industry, we can use them to estimate unrecorded intangibles. However, some industries, such as retail, have neither XRD nor XBRL tags for innovative investments. I apply textual analysis and machine learning tools to them and explain the methods in the next section.

3. Using NLP and W2V for firms with neither XRD nor relevant XBRD tags

3.1. **Developing Intangibles Dictionary**

I analyze 10-K texts using natural language processing (NLP) and machine learning tools such as Word2Vec (W2V) to identify firms that explain innovative investments. Loughran and McDonald (hereafter LM, 2011) is a seminal paper in finance NLP. They show that word lists developed by psychologists to evaluate sentiments in documents do not work well in finance research. It is because many English words have multiple meanings, and finance has many unique expressions, as other disciplines do. They develop dictionaries that reflect the tone of financial statements. See Loughran and McDonald (2016) for a review of textual analysis in accounting and finance. I thank them for making their dictionary and other data available for download from Software Repository for Accounting and Finance.

The LM method is a bag-of-words (BOW) approach to textual analysis because it regards a financial statement as a bag of all words in the document, regardless of how words are combined to explain the meanings of sentences. However, BOW has limitations in helping us understand financial statements because we need not only the list of single words but also how they are combined to form sentences. Thus I add W2V to the analysis of innovative investments.

W2V is a machine learning tool that learns the meaning of words in a set of documents, called corpus in computer science, by converting the relationships between words and phrases into a series of mathematical vectors. The method builds on the idea that words with similar meanings tend to appear with similar neighboring words (Harris, 1954). Mikolov et al. (2013) develop a method that trains a corpus to learn relationships between words and phrases. See Appendix B for technical details on W2V. Kai *et al.* (2021 a and b) apply this method to earnings call transcript data for textual analysis of corporate culture and the impact of Covid-19 on businesses. I apply W2V to 10-K texts for analyzing innovative investments.

Cao *et al.* (2021) built a machine-learning model to estimate stock prices, compare the machine's performance with analysts', and find that human analysts perform better than the model when a firm is complex with intangible assets. They analyze 10K texts only for sentiment analysis, not for estimating intangible assets. Their measure of intangible assets are based on Compustat data, and it is defined as the first principal component of four proxies: intangible assets minus goodwill divided by total assets, one minus the ratio of PP&E to total assets, organization capital scaled by assets, and knowledge capital scaled by assets, following Ewens *et al.* (2020). In contrast, I present machine learning of 10K texts as a valuable tool for analyzing unrecorded intangible assets.

For example, building a uniquely curated customer shopping experience gives retailers a competitive advantage in the digital economy. When I train a W2V model using randomly selected one thousand retail 10Ks as the corpus, the model identifies "product selection," "personalized, "merchandise assortments," "differentiation," and "memorable" as words related to "*unique*" in the first round. Iteration of this process builds a list of all words that describe unique shopping experiences in the retail industry. We can use this list to identify retailers that have invested in this area more extensively than their competitors.

All W2V models need seed words, as shown in the previous example that uses "unique" as a seed. Thus I first develop seeds for innovative investments using two primary sources. First, I collect seed words on intangibles from large accounting firms' websites advising on mergers and acquisitions involving intellectual property. For example, the categories of intangible assets include marketing-related, customer-related, contract-related, technology-related, and other unspecified intangible assets in KPMG's Castedello and Klingbeil (2009). As shown in Appendix A, unpatented technologies, databases, and many other seed words are from the website.

The second source of seed words is 10Ks of firms that invest heavily in unrecorded intangible assets. For example, retail is one of the most heavily affected industries that need innovation in the digital economy. Thus I start with 10Ks of retailers to build seed words and expand the search to other industries. Among the 8,549 retail 10Ks submitted to the SEC from 1994 – 2018, 693 firm-year (8.1%) mentions Amazon, a pioneer in digital transformation. The proportion increased sharply from 0% in 1994 to 25% in 2018, as shown in Table 5. Among the 693 10Ks, I randomly selected five percent as the sample and read them to build a list of words and phrases that describe unrecorded intangible investments.

For example, in the 10K filed on April 7, 2000, Nordstrom explains its e-commerce investment as follows: "*On November 1, 1999, the Company established a new subsidiary,.., to promote the rapid expansion of both its Internet commerce and catalog businesses. The Company contributed assets and certain liabilities associated with its Internet commerce and catalog businesses and $10 million in cash to the subsidiary… The Company has approximately 100 trademarks… all computer programs and software, and all rights and interests of Business in and to computer programs and software used primarily in connection with the Business, subject to any consents required for transfer of such computer programs and software.*" I take *Internet, trademark, computer,* and *software* from these sentences and add them to the list of seed words.

There are 103 words and phrases from the above two sources, and Figure 4 presents the wordcloud that shows how often they appear in the 8,549 retail 10Ks. It is an empirical question whether this 10K textual information on intangibles has any explanatory power of the numerical operating expenses data in Compustat based on GAAP. I use cross-sectional regressions to check the relationship and present the results in the following subsection.

3.2. Regressions of text data on expenses to estimate innovative investments

Table 4 presents summary statistics of retailers in Compustat for the fifty years from 1968 to 2018. On average, retailers spend 94 percent of their sales revenue on operation expenses (OExp). Among them, 71 percent is COGS, and the rest is XSGA. Advertising expenses (XAD) are 2 percent of sales on average, and XRD is negligible. The balance sheet data scaled by total assets (AT) show that the average common stock (CEQ) is 36 percent, inventory 28 percent, and goodwill and other recorded intangible assets are 8 percent of total assets.

When the digital economy started in the mid-1990s, the productivity and efficiency of retailers changed significantly. As shown in Figure 5, the labor productivity measured by sales revenue in 2010$ per employee increased from $152,000 in 1996 to $216,000 in 2018. The inventory-to-sales revenue ratio dropped from 13.4% to 9.6% during the same period. However, Compustat data do not show which retailer invested more in this innovation because the intangible investments are reported as part of XSGA and COGS in their income statements. Thus I apply the intangibles dictionary to each retailer's 10-K and test if the textual information explains the cross-sectional variation in operating expenses.

I analyze all 10-Ks retailers submitted to EDGAR during 1994 -2018 using the intangibles dictionary, iDic, developed in the previous section to calculate intangibles dictionary score (IDIS). IDIS is the total number of iDic words and phrases in 10-K divided by the total number of words multiplied by 100. The word counts are based on cleaned 10-K files after removing ASCII-encoded segments, tables, markup tags, and other irrelevant components from the EDGAR files. See Bodnaruk et al. (2015) for more details on the procedures to convert semistructured 10-K files from EDGAR to structured data for textual analysis.

I match the intangibles dictionary score, IDIS, based on the EDGAR text files with the SALE, COGS, and XSGA data in the Compustat database and present the summary statistics in Panel A of Table 6. Compustat defines the cost of goods sold (COGS) as all costs directly allocated by a company to production, such as material, labor, and overhead and selling, general and administrative expenses (XSGA) as all expenses not directly related to product production and incurred in the regular course of business pertaining to the securing of operating income. I define OExp, operating expenses, as the sum of COGS and XSGA in Compustat. Among the 8,549 retail EDGAR 10Ks, 7,237 firm-years have a matching central index key (CIK) in Compustat for the fiscal years 1995 - 2017. Among them, 6,504 observations have a SALE greater than zero, and there remain 6,439 firm-years after trimming outliers at the 1st and 99th percentiles of OExp/SALE.

The average number of words in the cleaned 10-K files is 40,718, and the average number of intangibles dictionary words and phrases is 224. 0.58% is the average proportion of intangibles words in the retail 10-Ks with a standard deviation of 0.28% and the 75th percentile of 0.72%. The average XSGA/SALE is 0.26, meaning that retailers, on average, spend 26% of their sales revenue on selling, general, and administrative expenses. The average OExp/SALE of 0.94 means that retailers pay 94% of their sales revenue for the cost of goods sold and selling, general and administrative expenses on average. Note that OExp includes investments in intangibles combined with administrative and other maintenance costs because the U.S. GAAP does not allow internally generated intangible assets to be capitalized. I use regressions to test whether IDIS can explain the cross-sectional variation in OExp/SALE and XSGA/SALE for separating intangible investments from other costs. I use the regression coefficient on IDIS to estimate unrecorded intangibles.

Panel B of Table 6 presents cross-sectional regressions of operating expenses normalized by sales revenue on IDIS and its variants. Model 1 in the table shows that the average retail firm-year that did not mention any words in the intangibles dictionary in their annual report spent 89 ~ 97% of their sales revenue on operating expenses, depending on the retail subsector. One percentage point increase in the intangibles word proportion, IDIS, results in a three percentage point increase in the operating expenses to the sales revenue ratio, and the difference is significant at the 1 percent level after controlling for the differences in eight subsectors.

Model 1: $OExp/SALE_{i,t} = \beta * IDIS_{i,t} + \text{industry dummies} + \varepsilon_{i,t,}$ (1)

where OExp is COGS plus XSGA. A measure of unrecorded intangible assets, $IDIS_{i,t}$, is defined as the total number of words and phrases from the Intangibles Dictionary in Firm i's 10-K submitted to the SEC in year t normalized by the total number of words in the 10-K and then multiplied by 100.

Model 2 shows that a change in the intangibles proportion from the previous year, $IDIS_t – IDIS_{t-1}$, has better explanatory power than $IDIS_t$. Retailers whose intangibles word proportion increased one percentage point from the prior year spent seven percent more operating expenses than other retailers after controlling for retail subsectors, and the difference is significant at the 1 percent level.

Model 2: $OExp/SALE_{i,t} = b * IDIS_{i,t-1} + c * (IDIS_{i,t} - IDIS_{i,t-1}) + \text{industry dummies} + \varepsilon_{i,t}$ (2)

Model 3 refines the explanatory variable further using I^P, an indicator variable with the value of one if $IDIS_t - IDIS_{t-1}$ is positive and zero otherwise. The idea is that the change in intangibles word proportion is meaningful for explaining cross-sectional differences in operating expenses only if the change is positive. The reason is that firms that start investing in intangibles present the investment in their annual report, which makes IDISt – IDISt-1 positive. Their $OExp/SALE_t$ is greater than other firms due to the intangible investments. When the firm completes intangible investments, its I_P will be zero, and it will have a similar OExp/SALE as its peers in the same retail subsector. Thus mentioning intangibles words less frequently than before does not mean operating expenses are lower than peers.

Consistent with this reasoning, the explanatory variable $(IDIS_t - IDIS_{t-1})*I_P$ in Model 3 shows the most significant result. The *t*-value is 11.34, and the adjusted-R^2 is over 95 percent. I also added year dummies, repeated the test, and found that year dummy variables are mostly insignificant, meaning that retailers' operating expenses are stable over time. The result for the model with year dummies is omitted to save space but is available from the author upon request.

Model 3: $OExp/SALE_{i,t} = \beta * (IDIS_{i,t} - IDIS_{i,t-1})\mathbf{*I^P}$ + industry dummies + $\varepsilon_{i,t}$ (3)

where, $\mathbf{I^P}$, is one if $IDIS_t - IDIS_{t-1}$ is positive and zero otherwise.

Model 4: $XSGA/SALE_{i,t} = \beta * (IDIS_{i,t} - IDIS_{i,t-1})\mathbf{*I^P}$ + industry dummies + $\varepsilon_{i,t}$ (4)

Model 4 is to show why OExp, not XSGA, should be used to estimate unrecorded intangibles. When comparing Models 3 and 4, we can see that the explanatory power of the increase in the intangibles dictionary score is much higher when we use OExp that includes both COGS and XSGA, than using only XSGA (adjusted-R^2: 95.54% vs. 62.80%, *t*-value: 11.34 vs. 8.90.

I argue that we should consider both XSGA and COGS when estimating unrecorded organization capital, and Models 3 and 4 in Table 6 of this paper present empirical results that support the argument. Note that prior research on intangibles uses only XSGA when estimating unrecorded organization capital and does not consider COGS. However, Ball et al. (2015) point out that COGS and XSGA are economically similar, and allocating expenses among them is determined arbitrarily by firms, not by GAAP. Crouzet and Eberly (2018) also show that COGS and XSGA are inconsistently defined across retailers, pointing out Walmart and Walgreens allocate wages and salaries to XSGA while Costco split them between XSGA and COGS, for example.

3.3. Estimated Unrecorded Intangible Assets in the Retailer Industry

The regression models presented in the previous section show that intangibles dictionary words explain the cross-sectional variation in operating expenses. The best is Model 3 that uses OExp/SALE, and 0.1264 is the highly significant coefficient for $(IDIS_t - IDIS_{t-1})*I_P$. Therefore, I use this result to define InvOExp, the intangible-investment proportion of operating expenses in Equation (5).

$$InvOExp_{i,t} \equiv 0.1264*(IDIS_{i,t} - IDIS_{i,t-1})* I_P*SALE_{i,t} \quad (5)$$

For example, if a retailer reported a SALE of \$3.5 billion for the fiscal year t and its IDIS increased from 0.44 in year t-1 to 0.99 in year t, I estimate that 0.1264*(0.99-0.44)*1*3.5 = \$243 million is the InvOExp in year t.

Panel A of Table 7 presents summary statistics of InvOExp scaled by SALE. 1.68% is the average unrecorded intangible investment of retailers as a proportion of the sales revenue in the 6,439 firm-years in the sample. The standard deviation is 2.9%, the median is 0.38%, and the 75th percentile is 2.19%. Over a quarter of the retail firm-years invested over two percent of their sales revenue in intangible assets and explained them in the financial statement texts during fiscal years 1995-2017. However, these investments do not appear as assets on their balance sheets due to accounting conservatism. The 1- 3 percent expensed investments have significant impacts on profits because retail is a low-margin industry.

I capitalize InvOExp, the intangible investment proportion of operating expenses, using a perpetual inventory method to estimate unrecorded intangibles. I assume that the depreciation rate of investments to develop organization capital is 0.2, and the depreciation rate for R&D and advertising expenses is 0.33, following Ewens et al. (2020). They estimate the parameters by incorporating publicly traded prices and exit prices from acquisitions, bankruptcies, and liquidations in a capitalization model. For example, the estimated InvOExp of $243 million in year t in the previous example accumulates to form organization capital, and it depreciates to (1-0.2)*243 = $194.4 million in year t+1. At the beginning of the sample period, I assume that retailers' initial stock of unrecorded intangibles was zero.

Summary statistics for the estimated unrecorded intangibles scaled by intangible-adjusted total assets, iAT, defined in Equation (6), are in Panel B of Table 7.

$$\text{iAT} \equiv \text{AT} + \text{Org} + \text{Know} + \text{Ad} - \text{GDWL} \tag{6}$$

where iAT is intangible-adjusted total assets, AT is total assets in Compustat, and Org is unrecorded organization capital estimated by capitalizing InvOExp. Know is unrecorded knowledge capital estimated by capitalizing XRD, the R&D expenditures in Compustat, Ad is the unrecorded intangibles estimated by capitalizing XAD, the advertising expenditures in Compustat, and GDWL is goodwill in Compustat. For example, if a retailer has neither GDWL nor Know but has AT of $2,131 million, Ad of $294 million, and Org of $ 243 million in year t, its iAT is 2,131 + 243 + 0 + 294 – 0 = $2,668 million in year t.

A firm accumulates its knowledge capital, Know, by spending on R&D. It is based on the idea that the outputs of R&D should be treated as capital rather than as intermediate input (Corrado et al., 2009). The same logic applies to Ad, capitalized advertising expenses. GDWL is the excess purchase price paid over the target's identifiable net assets' estimated fair value in business combinations.

Many retailers report advertising expenses, and the proportion is increasing. Among the 6,439 firm-years in the sample, 4,909 firm-years have XAD (76.24%), and the ratio was 57.01% in 1997 and 85.96% in 2017. In contrast, less than 5 percent of retail firm years (317 out of 6,439) report R&D expenditures, pointing to the importance of estimating InvOExp to sort out intangible investments from operating expenses.

Financial Accounting Standards Board (FASB) issued the Statement of Financial Accounting Standards (SFAS) 141 (Business Combinations) and SFAS 142 (Goodwill and Other Intangible Assets) to improve accounting standards on intangibles in 2001. According to these standards for mergers and acquisitions, acquirers must allocate the purchase prices they pay for targets to the tangible and identifiable intangible assets they acquire and the remainder to goodwill (FASB 2001a and 2007; FASB 2001b; Lim et al., 2020; Park 2019). FASB standards are now in Accounting Standards Codification (ASC). See ASC 805 for SFAS 141 and ASC 350-20-35 for SFAS 142.

I subtract GDWL when defining iAT in Equation 6 because of subjectivity in estimating goodwill's current fair value and goodwill impairments that are not backed by economic fundamentals (Ramanna and Watts, 2012; Chen et al., 2014). Park (2022) tests this theoretical reasoning of excluding goodwill empirically by comparing intangible adjusted book-to-market ratios with and without goodwill. She finds that the ratio excluding goodwill is superior to the measure including goodwill in explaining the cross-sectional variation in future stock returns.

As shown in Panel B of Table 7, an average retailer has 16.97% of its assets unrecorded in the balance sheet. The standard deviation is 11.62%, and the median is 14.64%. Among the unrecorded intangibles of retailers, the largest is unrecorded organization capital, Org, estimated by capitalizing InvOExp from the intangibles dictionary. Panel C of Table 7 presents Org/iAT by retail subsectors. The average unrecorded organization capital is above five percent of iAT in all retail subsectors. Grocery stores, on average, have a higher Org/iAT than restaurants (10.55% vs. 5.79%).

The average retailer's Org is 8.46% of its iAT, and the standard deviation is 6.92%, as shown in Panel B of Table 7. The second largest is Ad, estimated by capitalizing advertising expenses in Compustat. The average Ad of 8.12%, much larger than the median of 5.59%, implies that large retailers' advertising expenses make the Ad's distribution skewed to the right. The average Know is only 0.38% because most retailers do not have R&D expenses.

The low Know and high Org of retailers show the importance of separating intangible investments from operating expenses. This paper contributes to the literature by developing a novel textual analysis method and iDic, the intangibles dictionary. Unrecorded intangibles also affect profitability and investment measures.

Retail is a low-margin industry. The profit margin of retailers defined as aggregate income before extraordinary items, IB, divided by total SALE, is stable around 2-3 percent during the past fifty years. Therefore, whether to expense or capitalize the estimated intangible investments, InvOExp, significantly impacts retailers' measured profitability. Note that the average InvOExp/SALE in Table 7 is 1.68 percent.

Panel A of Table 8 presents the time series of estimated InvOExp of retailers compared to other intangible investments such as XAD and XRD from Compustat. Note that the estimated InvOExp is larger than XAD and IB in many years, implying that unrecorded intangible investments caused by accounting conservatism are significant compared to their profits and recorded intangibles. For example, during the fiscal year 2017, retailers generated a total sales revenue of $2.5 trillion while generating $92.5 billion as profits. They reported $37 billion as total advertising expenses, but the $22 billion investments in intangibles were combined with operating costs and expensed without being recorded separately.

I estimate Org, unrecorded organization capital, of retailers by capitalizing InvOExp and present the time series in Panel B of Table 8 compared to total assets, recorded other intangibles, goodwill, and capitalized advertising expenses and R&D expenditures. I define UI, total unrecorded intangibles, as the sum of Org, Know, and Ad.

$$\text{Unrecorded intangibles (UI)} \equiv \text{Org} + \text{Know} + \text{Ad} \tag{7}$$

As shown in Panel B of Table 8, Org is retailers' most significant component of unrecorded intangible assets, and its size is comparable to ROINTAN, recorded intangibles minus goodwill in Compustat. Total unrecorded intangibles, UI, of retailers are similar in size to recorded GDWL. For example, in the fiscal year 2017, retailers have $238 billion of goodwill recorded on their balance sheets. $239 billion is their total unrecorded intangible assets, which is estimated from their prior advertising, research and development, and other intangible investments based on this paper's textual analysis.

The unrecorded intangible investments of retailers have a significant impact on their profit margin. Panel C of Table 8 presents the time series of retailers' total IB to total SALE ratio (IB/SALE) compared with intangible-adjusted IB to SALE ratio (iIB/SALE). Equations

(8) and (9) define intangible investments (iInv) and intangible-adjusted income before extraordinary items (iIB).

$$iInv_{i,t} \equiv XAD_{i,t} + XRD_{i,t} + InvOExp_{i,t} \quad (8)$$

$$iIB_{i,t} \equiv IB_{i,t} + iInv_{i,t} - 0.33(Know_{i,t-1} + Ad_{i,t-1}) - 0.2Org_{i,t-1} \quad (9)$$

where $iInv_{i,t}$ is intangible investments of firm i during year t, $XAD_{i,t}$ is advertising expense from Compustat, $XRD_{i,t}$ is R&D expenditures from Compustat, $InvOExp_{i,t}$ is the investment proportion of operating expenses estimated based on the intangibles dictionary. $iIB_{i,t}$ is intangible-adjusted income before extraordinary items of firm i during year t by capitalizing intangible investments. I assume that the depreciation rates for Know, Ad, and Org of retailers are 0.33, 0.33, and 0.2, respectively, following Ewens et al. (2020).

As shown in Panel C of Table 8, intangible adjustments significantly impact retailers' profits. For example, the profit margin defined as total IB divided by total SALE of retailers in 1996 increased from 2.26% to 5.12% after adjusting for investments in intangible assets. Retailers invested \$16.3 – \$82.4 billion each year to develop intangibles during 1995 – 2017 when their IB was \$10.3 - \$92.5 billion per year. Therefore, I argue that we should reexamine the investment shortfall discussed in prior research using the estimated unrecorded intangible investments and discuss it in the next section.

3.4. Analyzing Investment Gap in the Digital Economy

In this section, I extend prior research on investment gaps using the estimated intangible investments. Hall (2014) analyzes the aggregate U.S. economic data and reports that investments in physical capital dropped sharply during the financial crisis of 2007-9, and the capital stock remains below trend even after the U.S. economic growth resumed. He examines investments in Intellectual Property separately from Plant, Equipment, and

Housing. He shows that Intellectual Property is the only component of the aggregate investment in the U.S. economy that remains on-trend. In contrast, all other areas show an investment gap from the trend.

Gutierrez and Phillippon (2017) and Alexander and Eberly (2018) build on this finding and analyze investment trends using firm-level data over time and across firms. They report that investments of U.S. firms fell relative to fundamentals such as cash flows and the market valuation of assets. They find that the investment gap started at the beginning of the 21st century, well before the financial crisis of 2007-9. Crouzet and Eberly (2018) build on these results and analyze the retail industry's investment gap.

They define the investment gap using tangible assets such as property, plant, and equipment, and I extend the research by adding estimated intangible investments to the analysis. I present intangible-adjusted investment gaps along with the tangible investment gap in Figure 3 and show that the investment gap reported in the literature using capital expenditures is no longer significant when investments include intangibles.

In Figure 6(a), I follow Crouzet and Eberly (2018) and regress the ratio of capital expenditure to property, plant, and equipment (CAPX/PPEGT) on Tobin's Q and the EBITDA to PPEGT ratio as shown in Equations (10) and (11). The investment gap presented in Figure 3(a) is the regression coefficient, δ_t, in Equation (10) for the sample period of 1995 – 2017. This regression model is based on the theory that firms increase capital spending when they have high operating profits and high market values compared to their assets' replacement costs. The intercept is the average CAPX/PPEGT in 1995, and the time dummies measure the investment gap for 1996 – 2017 after adjusting for profitability and valuation.

$$CAPX/PPEGT_{i,t} = \alpha_i + \delta_t + \beta_1 * Tobin'sQ_{i,t} + \beta_2 * EBITDA/PPEGT_{i,t} + \varepsilon_{i,t} \qquad (10)$$

Tobin's Q ≡ (Market value of equity +Book value of debt – Book value of current assets) / PPEGT

where Market value of equity = PRC * SHROUT from CRSP

Book value of debt = DLC plus DLTT from Compustat

Book value of current assets = ACT from Compustat (11)

The intangible investment gap presented in Figure 3(b) is the regression coefficient, d_t, in Equation (12). iAT and iInv are intangible-adjusted total assets and investments as defined in Equations (6) and (8), respectively. Total Q is defined as in Equation (13), following Peters and Taylor (2017).

$$iInv/iAT_{i,t} = a_i + d_t + b_1 * TotalQ_{i,t} + b_2 * iEBITDA/iAT_{i,t} + e_{i,t} \qquad (12)$$

Total Q ≡ (Market value of equity + iAT – Book value of equity) / iAT (13)

where Book value of equity = CEQ from Compustat

$$iEBITDA_{i,t} \equiv EBITDA_{i,t} + XAD_{i,t} + XRD_{i,t} + InvOExp_{i,t} \qquad (14)$$

Note that tangible investment shortfall in Fig6(a) is significant while intangible investment shortfall in Fig6(b) is not (*t*-statistic for 2017 - 1997: -9.26 tangible vs. 0.27 intangible). I also find that intangible adjustments affect the tangible investment gap as well. The intangible-adjusted tangible investment gap presented in Figure 6(c) is the regression coefficient, c_t, in Equation (15). The intangible-adjusted tangible investment shortfall in 2017 compared to 1997 in Fig3(c) is -3.28%, while the unadjusted shortfall in Fig 3(a) is -7.29%.

$$CAPX/iAT_{i,t} = a_i + c_t + b_1 * TotalQ_{i,t} + b_2 * iEBITDA/iAT_{i,t} + e_{i,t} \qquad (15)$$

These results confirm that unrecorded intangible investments are the main reason for the investment shortfall of retailers reported in the literature, and it is critical to adjust investment and profitability measures with intangibles when analyzing the retail industry.

When testing the relation between iDIS and OExp, I scaled the variables by SALE in the previous sections. As a robustness check, I use total assets (AT) to scale the variables and present the results in Table 9. As control variables, I add a revenue decrease dummy (D_SDrop) and an accounting loss dummy (D_Loss), following Enache and Srivastava (2018). They regress Sale/AT on XSGA/AT to estimate the investment component of XSGA by subtracting the maintenance proportion of operating expenses that vary with Sale. I add the textual information to the model and find a significant positive relation between ($IDIS_{i,t}$ – $IDIS_{i,t-1}$)* I_P and OExp/AT. Table 9 also confirms that OExp gives stronger results than XSGA as a measure of operating expenses, as in Table 6.

4. Analyzing patentless innovation using both XBRL and NLP

I used patents in previous sections to show that XBRL tags can explain innovation in no XRD firms, but there is much patentless innovation that increases firm values, and I analyze them using NLP of 10Ks. Many 10Ks explain how companies protect their intellectual property using trademarks, copyrights, patents, trade secrets, confidentiality agreements, and other internal policies and procedures.

Hoberg and Maksimovic (2015) present that many companies discuss proprietary information in their financial statements in the following contexts: the risk of potential damages when proprietary information is revealed or contracts with employees that forbid leaking proprietary information. They identify firms with proprietary information risks as those mentioning protect or safeguard proprietary information, trade secrets, or confidential information.

I develop an intellectual property dictionary (iPDic) using the following seed words: trademark, copyright, patent, trade secret, confidential, protect, safeguard, proprietary, and intellectual property. Then I search for the words in the 10Ks of financial firms and use the iPDic counts scaled by the total number of words to test if previous expenses on information technology are related to the use of iPDic words in 10Ks in the future.

Panel A of Table 10 presents summary statistics of the variables in the regressions to test the relation for the fiscal years 2009 – 2021, and there are 11,746 firm-years. The average COM, CIT, and ITDP scaled by Sale is 1.6 ~ 4.13% and the average iPDic counts scaled by the total number of words (NW) in 10K is 0.06%. Panel B of Table 10 presents regressions of iPDic/NW on the previous three years' expenses scaled by the total sales for the same period after controlling for the subsector effect. Brokers and Dealers (2-digit SIC: 62), on average, mentions intellectual property words more often, and Depository institutions (2-digit SIC:60) use those less frequently than others.

Note that the Compustat variable, XCOM, does not have explanatory power for the use of intellectual property words but expenditures based on XBRL tags, COM, CIT, ITDP, and COMITDP, show significant results. The communication and information technology tag (CIT) shows the impact on intellectual property with a coefficient of 0.27 and a *t*-statistic of 2.63. One percentage point increase in CIT to Sale ratio during the past three years resulted in a 0.27 percentage point increase in the use of intellectual property words in 10Ks.

I used financial firms to explain how to apply XBRL tags and textual analysis to evaluate the impact of prior expenditures on intellectual property in this section, but this methodology can be applied to many other industries. There is significant variability across industries in how they record intangible investments (Eisfeldt et al., 2022). In some sectors, we need both XBRL tags and textual data to analyze unrecorded intangibles, and we can develop a unique dictionary for each industry using W2V.

For example, Food (2-digit SIC: 20) is another industry with many valuable patents without R&D expenditures, as shown in Table 1. When I applied W2V to 1,000 randomly selected 10-Ks in the food industry with iDic as seed words, the machine learning model identified the following new words for intangibles specific to the sector; *cloud-computing, colorful, contemporary, distinctive, flavor_profiles, flavors, formulation, healthy_lifestyle, improvements_discoveries, niche, noncompetition, nondisclosure, confidentiality, nutritious_eating, packaging_innovations, proprietary_farming, proven_techniques, providing_customizable, quality_consistency, quality_taste, quickly_economically, recombinant, reformulate_exising, unpatented, reputation_brand, retain_talented, shopping_platform, skilled, sophisticated_computer, sponsored_research, successfully_introduce, tasting_opportunites, and technology_infrastructure, unique_flavor*

It is beyond the scope of this paper to develop intangible dictionaries for all industries and identify relevant XBRL tags for them. However, the XBRL and NLP tools presented in this paper for financial firms and retailers can be applied to all other industries.

5. Conclusion

Financial economists have used numerical information in financial statements summarized by Compustat for asset pricing since the mid-1960s. However, we are not fully utilizing the benefits of the digitalized financial statements the SEC has started introducing since the mid-1990s. This paper uses XBRL tags and text data to analyze innovative investments that are not recorded as assets but generate long-term benefits. XBRL tags show more granular information than Compustat, and financial statement texts reveal more details than numbers present.

I analyze innovation in the retail industry using textual analysis to estimate unrecorded intangible assets. I use the finance industry as an example to show how XBRL tags on information technology expenses in prior years explain patent values in the future. I also present cases needing both XBRL tags and textual analysis to analyze innovative investments.

Prior research analyzes innovative investments using R&D expenditures, but this paper shows that many firms innovate without reporting R&D expenditures. I show that we can use XBRL tags and financial statement texts to identify those innovators and estimate their unrecorded intangible assets to improve valuation models.

REFERENCES

Alexander, L. & J. Eberly. (2018). Investment hollowing out. *IMF Economic Review*, 66, 5-30.

Autor, D., D. Dorn, L.F. Katz, Ch. Patterson, & J. V. Reenen. (2017). Concentrating on the fall of the labor share. *American Economic Review: Papers & Proceedings*, 107(5), 180-185.

Autor, D., D. Dorn, L.F. Katz, Ch. Patterson, & J. V. Reenen. (2019). The fall of the labor share and the rise of superstar firms. *Quarterly Journal of Economics*, forthcoming

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. V. (2015). Deflating profitability. *Journal of Financial Economics*, 117, 225-248.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. V. (2020). Earnings, retained earnings, and book-to-market in the cross-section of expected returns. *Journal of Financial Economics*, 135(1), 231-254.

Barefoot, K., Curtis, D., Jolliff, W., Nicholson, J. R. & Omohundro, R. (2018). Defining and measuring the digital economy. Bureau of Economic Analysis, U. S. Department of Commerce.

Beaver, W., & Ryan, S. (2005). Conditional and unconditional conservatism: Concepts and modeling. *Review of Accounting Studies,* 10, 269-309.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, *50*(4), 623–646.

Branstetter, L. & Sichel, D. (2017). The case for an American productivity revival. Peterson Institute for International Economics (PIIE) Policy Brief.

Cao, S., Jiang, W., Wang, J. L. & Yang, B. (2021). From Man vs. Machine to Man + Machine: The Art and AI of Stock Analysis, NBER woring paper 28800

Castedello, M., & Klingbeil, C. (2009). Intangible assets and goodwill in the context of business combinations: An industry study. KPMG.

Chychyla, R. & Kogan, A. (2015). Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings. *Journal of Information Systems*, 29(1), 37-72.

Chen, W., Shroff, P. K., & Zhang, I. (2014). Fair value accounting: Consequences of booking market-driven goodwill impairment. Working Paper, University of Minnesota.

Corrado, C., Haskel, J., Jona-Lasinio, C. & Iommi, M. (2012). Intangible Capital and Growth in Advanced Economies: Measurement Methods and Comparative Results. Imperial College London Business School Working Paper.

Corrado, C. Hulten, C. & Sichel, D. (2009). Intangible capital and U. S. economic growth. *The Review of Income and Wealth*, 55(3), 661-685.

Corrado, C. & Slifman, L. (1999). A decomposition of productivity and costs. *American Economic Review*, 89, 328–32, 1999.

Crouzet, N. & J. Eberly. (2018). Investment, rates and rents: Intangibles, investment, and efficiency. *AEA Papers and Proceedings*, 108, 426-431.

Crouzet, N. & J. Eberly. (2019). Understanding weak capital investment: the role of market concentration and intangibles. Working paper 25869, National Bureau of Economic Research.

Dong, Y., O. Z. Li, Y. Lin, and C. Ni. (2016). Does information-processing cost affect firm-specific information acquisition? Evidence from XBRL adoption. Journal of Financial and Quantitative Analysis 51 (2): 435–462.

Eisfeldt, A. L. & Papanikolaou, D. (2013). Organization Capital and the Cross-Section of Expected Returns. *Journal of Finance*, 68(4), 1365 - 1406.

Elsby, M., B. Hobijn, & A. Sahin. (2013). The Decline of the U.S. Labor Share. *Brookings Papers on Economic Activity*, 1-42.

Enache, L. & A. Srivastava. (2018). Should Intangible Investments Be Reported Separately or Comingled with Operating Expenses? New Evidence. *Management Science*, 64(7), 3446 - 3468.

Ewens, M., R. Peters, & S. Wang. (2020). Measuring Intangible Capital with Market Prices. Working Paper, California Institute of Technology.

Financial Accounting Standards Board (1974). Statement of Financial Accounting Standards No. 2 Accounting for Research and Development Costs. Financial Accounting Standards Board (FASB), Norwalk, CT.

Financial Accounting Standards Board (2001a). Statement of Financial Accounting Standards (SFAS) No. 141 Business Combinations. FASB, Norwalk, CT.

Financial Accounting Standards Board (2001b). Statement of Financial Accounting Standards (SFAS) No. 142 Goodwill and Other Intangible Assets. FASB, Norwalk, CT.

Financial Accounting Standards Board (2007). Statement of Financial Accounting Standards (SFAS) No. 141 Business Combinations (Revised). FASB, Norwalk, CT.

Gutierrez, G. & T. Philippon. (2017). Investmentless Growth: An Empirical Investigation, Brookings Papers on Economic Activity

Hall, R. (2014). Quantifying the Lasting Harm to the U.S. Economy from the Financial Crisis. National Bureau of Economic Research (NBER) Macroeconomics Annual, 29(1), 71-218.

Hall, B. H. & Lerner, J. (2009). The financing of R&D and innovation. Working paper. National Bureau of Economic Research.

Harris, Z. S., 1954. Distributional structure. *Word* 10:146–162.

Hoberg, G. & Maksimovic, V. (2015). Redefining financial constraints: A text-based analysis. *Review of Financial Studies*, 28(5), 1312-1352.

Jolliff, W. & Nicholson, J. R. (2019). Measuring the digital economy: an update incorporating data from the 2018 comprehensive update of the industry economic accounts. Bureau of Economic Analysis, U. S. Department of Commerce.

Kogan, L., Papanikolaou, D., Seru, A. and Stoffman, N., 2017. Technological innovation, resource allocation, and growth. Quarterly Journal of Economics, 132(2), 665-712.

Koh, D., Santaeulalia-Llopis, R. & Zheng, Y. 2020. Labor share decline and intellectual property products capital. *Econometrica*, forthcoming.

Kothari, S. P., Laguerre, T. E., & Leone, A. J. (2002). Capitalization versus expensing: Evidence on the uncertainty of future earnings from capital expenditures versus R&D outlays. *Review of Accounting Studies*, 7, 355–382.

Lev, B., & Zarowin, P. (1999). The boundaries of financial reporting and how to extend them. *Journal of Accounting Research*, 37, 353–385.

Lev, B., & Radhakrishnan, S. (2005). The valuation of organization capital, in Corrado, C., Haltiwanger, J., & Sichel, D. eds.: Measuring Capital in the New Economy, National Bureau of Economic Research, Inc, Cambridge, MA.

Li, W. C. Y. (2012). Depreciation of business R&D capital. US Bureau of Economic Analysis/ National Science Foundation R&D Satellite Account Paper US Department of Commerce.

Li, W. C. Y. & Hall, B. W. (2016). Depreciation of business R&D capital. U.S. Bureau of Economic Analysis/ University of California at Berkeley and NBER.

Lim, S. C., Macias, A. J. & Moeller, T. (2020). Intangible assets and capital structure. *Journal of Banking and Finance*, forthcoming.

Loughran, T., & B. McDonald, (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66, 35–65.

Loughran, T., & B. McDonald, (2016). Textual analysis in accounting and finance: A Survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Myers, S., & N. Majluf, (1984). Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics*, 13, 187-221.

Michelacci, C. & Quadrini, V. (2009). Financial markets and wages. *The Review of Economic Studies*, 76 (2), 795–827.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Nakamura, L. (2001). What is the U.S. gross investment in intangibles? At least one trillion dollars a year. Federal Reserve Bank of Philadelphia Working Paper No. 01-15.

Nakamura, L. (2003). A trillion dollars a year in intangible investment and the new economy. In *Intangible assets*, J. Hand and B. Lev (eds.). Oxford University Press.

Park, H. (2019). Intangible assets and the book-to-market effect. *European Financial Management*, 25, 207-236.

Park, H. (2022). An intangible-adjusted book-to-market ratio still predicts stock returns, *Critical Finance Review*, 11(2), 265 – 297.

Paul, B., & Durbin, P. (2016). Revisiting accounting for software development costs: An ideal first step. *Point of View*, PricewaterhouseCoopers (PwC), October.

Penman, S. H., & Zhang, X. (2002). Accounting conservatism, the quality of earnings and stock returns. *The Accounting Review*, 77, 237-264.

Peters, R. H. & Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123, 251-272.

Ramanna, K., & Watts, R. (2012). Evidence on the use of unverifiable estimates in required goodwill impairment. *Review of Accounting Studies,* 17, 749–780.

Rassier, D. G. (2014). Treatment of research and development in economic accounts and in business accounts. *BEA Briefings*, March 1-8., Bureau of Economic Analysis (BEA), U.S. Department of Commerce.

Appendix A. Categories of Intangibles and Words and Phrases Used to Describe Them

Source: Castedello, M., & Klingbeil, C. (2009). Intangible assets and goodwill in the context of business combinations: An industry study. KPMG.

Illustrative examples for intangible assets according to IFRS3 and SFAS141 (p.21)

Technology related:
Patented technologies
Computer software and mask works
Unpatented technologies
Databases, including title plants
Trade secrets such as secret formulas, processes and recipes

Contract related:
Licenses, royalties, standstill agreements
Advertising, construction management, service, delivery and supply contracts
Lease agreements (independently of whether the acquiree is the lessee or the lessor)
Construction permits
Franchise agreements
Operating and broadcasting rights
Servicing contracts, such as mortgage servicing contracts
Use rights, such as drilling, water, air, timber cutting and route authorizations
Employment contracts

Customer related:
Customer lists
Order or production backlog
Customer contracts and related customer relationships
Non-contractual customer relationships

Marketing related:
Trademarks, trade names, service names, collective marks, certification marks
Trade dress (unique color, shape or package design)
Newspaper mastheads
Internet domain names
Non-competition agreements

Art related:
Plays, operas and ballets
Books, magazines, newspapers and other literary works
Musical works such as compositions, song lyrics and advertising jingles
Pictures and photographs
Video and audiovisual material,
including motion pictures or films, music videos and television programs

Appendix B. Machine Learning of Financial Statements Using Word2Vec (W2V)

This appendix is to present technical details on how to train a W2V model using financial statements such as 10Ks as a corpus. The first step is to download the raw 10K files using the SEC/EDGAR full index directory, https://www.sec.gov/Archives/edgar/full-index/. Then I clean the raw data to remove exhibits, HML and XBRL markup tags, and other non-text items that are irrelevant to textual analysis. I use the regex version 2021.8.3 and beautifulsoup 4 in Python 3.8 for the cleaning process.

The next step is to convert all alphabet in the cleaned 10Ks to lower case to facilitate word search, delete numbers and special characters irrelevant to textual analysis, and to use the Phraser and Phrases modules in the genism library to form multiword ngrams that are essential to learn how words are combined to deliver meanings. The modules automatically detect phrases longer than one word using collocation statistics. For example, the Phrases module makes "confidentiality_agreement' out of "confidentiality" and "agreement" and add the bigram, the combined word using the underscore symbol _, to the corpus. W2V treats an ngram concatenated with underscore like a single word.

As usual in most textual analysis projects, I remove stopwords such as "are," for example, that are used in most documents but do not add value in analyzing the meaning of words and phrases. See Appendix C for the list of stopwords used in this paper. For example, *"we depend on sophisticated information technology systems and a cyber attack or other breach of these systems could have a material adverse effect on our operations"* is converted to *"depend sophisticated information technology systems cyber attack breach systems material adverse effect operations"* after removing the stop words.

Stemming such as converting "competition" and "compete" to "compet", for example, is also used in many textual analysis projects. However, I find that the cost of lost information outweighs the benefit of reduced dimension when stemming 10Ks by comparing the results with and without stemming. It is because currently available stemming tools have been developed mostly outside of finance and thus do not consider the characteristics of terms used frequently in financial statements. Thus, the results reported in this paper are from the 10K corpus without stemming.

After removing the stopwords, I train the remaining words and ngrams using the W2V model in the genism library version 4.1.2. The model relies on word embedding that represents the meaning of a word using a numeric vector so that we can use vector arithmetic to measure the relationship between words. Mathematically W2V uses the cosine similarity between two word-vectors to measure how close the two words are.

For example, when we use vector arithmetic with how often [digital, ecommerce, inflation, operating_efficiencies, analytics, guest_experience] appear close to (online, supply_chain, marketing) in twenty 10-Ks of department stores to examine the relationships among the three words, we first need the following three vectors.

online = [25, 27, 3, 4, 8, 6], supply_chain = [10, 6, 18, 15, 12, 1], marketing = [12, 16, 0, 5, 17, 16]. The 25 and 27 in the online vector mean “digital” and “ecommerce” appear close to “online” 25 and 27 times, respectively. The window size in a W2V model defines what is regarded as appearing close. The window size of 3, for example, means three or fewer between two words are regarded as being close and thus count toward the vectors. The definition of the cosine similarity between vectors is as follows.

Cosine similarity between vectors A and B $\mathbf{= \Sigma A_i B_i / (\sqrt{\Sigma A_i^2}\sqrt{\Sigma B_i^2})}$

Cosine similarity between online and supply_chain = 628/(sqrt(1479)*sqrt(830)) = 0.57

Cosine similarity between online and maketing = 984/(sqrt(1479)*sqrt(970)) = 0.82

The higher cosine similarity of 0.82 vs. 0.57 means that online is more closely related to marketing than to supply_chain in the 10K texts of the department stores.

The above example shows how we can use vectors to quantify the relationship between any pair of words and phrases as well as what kind of challenges we face when applying this method to financial statements. We had only six components in the above vectors meaning that we represented the word “online” using only six words, but tens of thousands words appear in 10Ks and the dimension grows exponentially with phrases that are combinations of words.

The example also shows a clue to reduce the dimension to make this vectorization method practical. When we use simple counting of words as in the previous example when forming vectors, the implicit assumption is the index words [digital, ecommerce, inflation, operating_efficiencies, analytics, guest_experience] are orthogonal, meaning no relationship between “digital” and “ecommerce”, for example, which is not true, leading to unnecessarily many zeros or smaller numbers with higher-dimensional vectors. We can reduce dimension and unnecessary zeros significantly by using combinations instead of all words and ngrams in the corpus.

Mikolov et al. (2013) is a seminar paper addressing the issue by word embedding and this model is called W2V. They applied backpropagation, a training algorithm common in neural networks, to make parameters in the network adjusted and become an effective vector representation of a word when the learning is complete after iterations through the corpus. The neural network of word embedding works like concatenated regressions where hidden layers receiving output from the previous layer as an input and feeding the output forward to the next layer. The weight matrix randomly selected initially for the vectors continually improves as a backpropagation algorithm in a feed-forward neural network learns from mistakes and make adjustments. The learning is complete after the neural network is adept at the task after passing through the entire corpus iteratively and the result is a final vector representation for the trained corpus. Kai et al. (2021a and b) applies this method to earnings call transcripts to analyze corporate culture and to examine the impact of Covid-19 on businesses and their responses. When training the model with 10Ks for this paper, I set the window size to 5, the number of iterations to 30, and the minimum word count in the corpus to be considered to 3.

Appendix C. The list of 588 stopwords used to prepare 10K corpuses for W2V

Stopwords = [a, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, april, are, aren't, around, as, a's, aside, ask, asking, associated, at, august, available, away, awfully, b, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c, came, can, cannot, cant, can't, cause, causes, certain, certainly, changes, clearly, c'mon, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, c's, currently, d, december, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, done, don't, down, downwards, during, e, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, f, far, february, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, g, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, h, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, hello, help, hence, her, here, hereafter, hereby, herein, hereof, here's, hereupon, hers, herself, hereunder, he's, hi, him, himself, his, hither, hopefully, how, howbeit, however, I, i'd, ie, if, ignored, ii, iii, i'll, i'm, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, item, it'll, its, it's, itself, iv, i've, ix, j, january, july, june, just, k, keep, keeps, kept, know, known, knows, l, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, m, mainly, many, march, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, n, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, next, nine, no, nobody, non, none, nor, normally, not, nothing, novel, november, now, nowhere, o, obviously, october, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, p, page, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, q, que, quite, qv, r, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, s, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, september, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that’s, that's, the, their, theirs, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, thereof, theres, there's, thereto, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, too, took, toward, towards, tried, tries, truly, try, trying, t's, twice, two, u, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, v, various, very, vi, via, vii, viii, viz, vs, w, want, wants, was, wasn't, way, we, we'd, welcome, well, we'll, went, were, we're, weren't, we've, what, whatever, what's, when, whence, whenever, where, whereafter, whereas, whereby, wherein, where's, whereupon, wherever, whether, which, while, whither, who, whoever, whole, whom, who's, whose, why, will, willing, wish, with, within, without, wonder, won't, would, wouldn't, x, y, yes, yet, you, you'd, you'll, your, you're, yours, yourself, yourselves, you've, z, zero

TABLE 1. Patents Granted to Firms without R&D Expenses: 1974 – 2020

This table presents the number of patents and their total market value deflated to 1982 dollars as in Kogan et al. (2017) during the sample period from 1974 to 2020 for the Compustat firms that do not have R&D expenditures (XRD). I classify the patents by the two-digit SIC codes of the companies and present the total number of citations these patents have received as of December 31, 2020.

SIC	Industry	No of patents	Market value (1982 $ million)	Cites
10	Metal mining	161	8323	1322
13	Oil and gas	1111	26281	19326
20	Food	2152	201198	42722
21	Tobacco products	357	12636	5506
22	Textile mill	236	1073	2963
23	Apparel	326	3316	1673
26	Paper products	1138	9438	13352
27	Printing	707	4409	11417
28	Chemicals	1778	6633	32046
29	Petroleum refining	204	8584	2651
30	Rubber and plastic	4659	88407	34077
33	Primary metal industries	651	7095	5913
34	Fabricated metal	1126	11159	15664
35	Machinery	1993	31224	32495
36	Electronic equipment	1453	4107	23683
37	Transportation equipment	6677	14915	95841
38	Measuring instruments	860	2596	22911
39	Miscellaneous manufacturing	419	2694	9935
40	Railroad	133	4867	1220
42	Transportation	648	42137	8015
45	Air transportation	290	18434	2235
48	Telecommunication	19175	752027	149875
49	Electric services	2664	21224	14231
57	Home furniture	238	3581	5975
58	Restaurants	64	6460	580
59	Retail	319	22039	2300
60	Depository institutions	6510	757288	33217
61	Non-depository credit institutions	3450	101924	22592
62	Brokers and dealers	1338	83506	7807
63	Insurance	2177	32443	45594
73	Business services	1524	19455	27832
75	Auto repair	139	1104	5144
79	Recreation services	98	2374	3133
80	Health services	352	5767	2209
87	Engineering	1107	8492	7347

TABLE 2. XBRL Tags for Innovative Investments in the Finance Industry

This table presents the summary statistics of the three XBRL tags identified to measure innovative investments in the finance industry for the sample period of 2009 – 2021, and the unit is US$ million: Communication (COM), CommunicationsAndInformationTechnology (CIT), and InformationTechnologyAndDataProcessing (ITDP). I define COMITDP as the sum of COM, CIT, and ITDP and include it in this table because some firms report more than one of the three tags for the same year.

Tag	Number	Mean	Standard Deviation	Percentiles		
				25th	Median	75th
COM	769	31.6	153	0.7	1.8	5.8
CIT	1,026	324.3	1,160	3.4	17.7	111.0
ITDP	2,903	31.3	203	1.2	2.6	7.3
COMITDP	4,132	108.4	632	1.4	3.7	16.0

TABLE 3. Regressing Patent Values on Prior Expenses on Information Technology

This table is to test whether previous expenditures on information technology and data processing of financial firms explain the value of patents they receive in the future. In the regressions, I use the estimated market value of patents as the dependent variable and the values for the three years prior to the patent grant year for the following XBRL tags as explanatory variables: Communication (COM), CommunicationsAndInformationTechnology (CIT), and InformationTechnologyAndDataProcessing (ITDP). I also test COMITDP, the sum of COM, CIT, and ITDP, and XCOM from Compustat that is comparable to COMITDP because some firms report more than one of the three tags for the same year. The numbers in parentheses are *t*-statistics and *** denotes the coefficient is significant at the 1 percent level.

Explanatory variable	Model 1	Model 2	Model 3	Model 4	Model 5
COM	9.66 (6.95)***				
CIT		1.96 (9.70)***			
ITDP			4.88 (10.06)***		
COMITDP				2.35 (13.05)***	
XCOM from Compustat					2.04 (11.15)***
Adj-R^2(%)	62.02	68.42	78.16	66.07	36.79

TABLE 4. Summary Statistics of Compustat Data for Retailers: 1968 – 2018

This table presents summary statistics of retailers in the Compustat database and the sample period is fiscal years 1968 – 2018. Operating expenses (OExp) is defined as the sum of cost of goods sold (COGS) and selling, general and administrative expenses (XSGA). There are 22,536 firm-years and the means in the table are after trimming at the 1st and the 99th percentiles to remove outliers. The balance sheet items are scaled by total assets (AT) and income statement variables are scaled by sales revenue (SALE).

Variable	Mean	Standard Deviation	Percentiles		
			25th	Median	75th
Accounting data scaled by SALE					
COGS	0.7105	0.1478	0.6267	0.7202	0.8075
XSGA	0.2453	0.1479	0.1487	0.2284	0.3117
OExp	0.9389	0.0977	0.8985	0.9353	0.9664
XAD	0.0195	0.0269	0.0000	0.0110	0.0297
XRD	0.0004	0.0042	0.0000	0.0000	0.0000
EBITDA	0.0611	0.0977	0.0336	0.0647	0.1015
IB	0.0014	0.1127	0.0009	0.0188	0.0393
CAPX	0.0516	0.0886	0.0164	0.0303	0.0555
PPEGT	0.3480	0.3782	0.1475	0.2533	0.4440
AT	0.6651	1.3870	0.3703	0.5053	0.6871
Balance sheet items scaled by AT					
INVT	0.2801	0.2081	0.0879	0.2606	0.4260
ACT	0.5046	0.2377	0.3106	0.5271	0.6958
DLC	0.0722	0.2133	0.0038	0.0221	0.0721
DLTT	0.2264	0.2931	0.0507	0.1800	0.3216
CEQ	0.3619	0.7324	0.2693	0.4260	0.5702
GDWL	0.0415	0.0995	0.0000	0.0000	0.0188
ROINTAN	0.0341	0.0812	0.0000	0.0000	0.0260

Table 5. Companies Mentioning the Largest e-commerce Firm in Financial Statements

This table presents the number and proportion of retailers that mention Amazon in their 10-K annual financial statements filed with the SEC compared to all firms in the EDGAR filing system. Among the 1,028,674 documents submitted to the SEC EDGAR during 1994 – 2018, 181,426 are 10-K annual reports. Among the 181,426 10-Ks, 8,549 files belong to the retail industry based on the SIC codes in the 10-Ks.

Filing Year	All Firms			Retailers		
	Total	Mentioning AMZN	Proportion	Total	Mentioning AMZN	Proportion
1994	1910	7	0.37%	130	0	0.00%
1995	2213	9	0.41%	150	0	0.00%
1996	4293	7	0.16%	227	0	0.00%
1997	6640	17	0.26%	364	2	0.55%
1998	6861	33	0.48%	372	4	1.08%
1999	6716	74	1.10%	349	10	2.87%
2000	6578	125	1.90%	341	21	6.16%
2001	6225	90	1.45%	305	16	5.25%
2002	6670	92	1.38%	375	17	4.53%
2003	8433	111	1.32%	395	17	4.30%
2004	8524	95	1.11%	368	25	6.79%
2005	8997	117	1.30%	353	27	7.65%
2006	8821	118	1.34%	350	28	8.00%
2007	8524	122	1.43%	358	25	6.98%
2008	8641	148	1.71%	366	23	6.28%
2009	9785	185	1.89%	422	29	6.87%
2010	9096	201	2.21%	429	29	6.76%
2011	8754	231	2.64%	416	32	7.69%
2012	8333	280	3.36%	394	43	10.91%
2013	7999	342	4.28%	383	42	10.97%
2014	7955	395	4.97%	377	49	13.00%
2015	7845	438	5.58%	351	50	14.25%
2016	7452	484	6.49%	336	57	16.96%
2017	7157	518	7.24%	326	69	21.17%
2018	7004	626	8.94%	312	78	25.00%

TABLE 6. Textual Analysis of Intangible Investments in the Retail Sector

This table presents a textual analysis of retail 10K annual reports matched with numerical information in Compustat. Panel A presents summary statistics, and Panel B shows regressions to test the relation between intangibles dictionary score (iDIS) and the operating expenses to sales ratio (OExp/SALE). I^P is a dummy variable that is one if the iDIS has increased from the previous fiscal year and zero otherwise. The numbers in parentheses are *t*-statistics, and *** means the statistical significance at the one percent level.

Panel A: Summary statistics

Variable	Mean	Standard Deviation	Percentiles		
			25th	Median	75th
OExp/SALE	0.94	0.21	0.88	0.93	0.96
XSGA/SALE	0.26	0.22	0.13	0.23	0.32
Number of words	40,718	30,403	22,966	33,476	48,656
Intangibles dictionary counts	224	170	107	185	294
iDIS	0.58	0.28	0.39	0.53	0.72

Panel B: Regressions

Models	Model 1	Model 2	Model 3	Model 4
(Dependent Variable)	(OExp/SALE)	(OExp/SALE)	(OExp/SALE)	(XSGA/SALE)
Intercept				
iDIS$_t$	0.03 (2.76)***			
iDIS$_{t-1}$		-0.01 (-0.34)		
iDIS$_t$ - iDIS$_{t-1}$		0.07 (6.68)***		
(iDIS$_t$ - iDIS$_{t-1}$)*I^P			0.13 (11.34)***	0.10 (8.90)***
Hardware (52)	0.92	0.93	0.92	0.26
Department stores (53)	0.92	0.93	0.92	0.24
Grocery (54)	0.93	0.94	0.93	0.22
Auto and gas (55)	0.92	0.93	0.92	0.20
Clothing (56)	0.89	0.91	0.89	0.29
Furniture/appliance/computer (57)	0.95	0.96	0.94	0.31
Restaurants (58)	0.90	0.91	0.90	0.13
Miscellaneous (59)	0.97	0.98	0.97	0.32
Adj-R^2(%)	95.46	95.50	95.54	62.80

TABLE 7. Unrecorded Intangible Assets in the Retail Industry

Panel A of this table presents the descriptive statistics for the investment proportion of operating expenses, InvOExp defined in Equation (5) for the 6,439 retail firm-years in fiscal years 1995 – 2017. Unrecorded organization capital, Org, in Panel B, is from capitalizing InvOExp by assuming that the depreciation rate is 0.2. The unrecorded advising capital, Ad, is from capitalizing XAD in Compustat and the unrecorded knowledge capital, Know, is from capitalizing XRD in Compustat. I assume that the depreciation rate for XRD and XAD is 0.33, following Ewens et al. (2020). Unrecorded intangibles, UI, is the sum of Org, Ad, and Know. Intangible-adjusted total assets, iAT, is defined in Equation (7) as the sum of total assets in Compustat (AT), Org, Ad, and Know minus goodwill in Compustat (GDWL). Panel B presents descriptive statistics of UI, Org, Ad, and Know normalized by iAT. Panel C compares unrecorded organization capital scaled by intangible-adjusted total assets by retail subsectors.

Panel A: Summary statistics of the investment proportion of operating expenses normalized by sales revenue by retail subsector

InvOExp/SALE	Number of Firm-years	Mean	Standard Deviation	Percentiles		
				25th	Median	75th
All retail	6439	0.0168	0.0290	0.0000	0.0038	0.0219
Hardware	207	0.0146	0.0244	0.0000	0.0023	0.0170
Department stores	556	0.0151	0.0257	0.0000	0.0032	0.0215
Grocery	502	0.0141	0.0209	0.0000	0.0039	0.0205
Auto and gas	467	0.0143	0.0233	0.0000	0.0028	0.0179
Clothing	936	0.0164	0.0282	0.0000	0.0039	0.0214
Furniture, appliance, and computer	455	0.0202	0.0359	0.0000	0.0057	0.0239
Restaurants	1557	0.0136	0.0230	0.0000	0.0032	0.0165
Miscellaneous	1759	0.0212	0.0356	0.0000	0.0047	0.0288

Panel B: Unrecorded intangible assets of retailers scaled by intangible-adjusted total assets

Unrecorded intangibles scaled by iAT	Mean	Standard Deviation	Percentiles		
			25th	Median	75th
Unrecorded intangibles	0.1697	0.1162	0.0894	0.1464	0.2228
Know	0.0038	0.0277	0.0000	0.0000	0.0000
Ad	0.0812	0.0930	0.0077	0.0559	0.1184
Org	0.0846	0.0692	0.0408	0.0685	0.1083

Panel C: Unrecorded organization capital scaled by intangible-adjusted total assets by retail subsectors

Org scaled by iAT	Mean	Standard Deviation	Percentiles		
			25th	Median	75th
Hardware	0.0888	0.0895	0.0335	0.0632	0.1160
Department stores	0.0840	0.0603	0.0478	0.0672	0.1012
Grocery	0.1055	0.0565	0.0648	0.0938	0.1325
Auto and gas	0.0847	0.0573	0.0406	0.0767	0.1143
Clothing	0.0865	0.0534	0.0495	0.0783	0.1104
Furniture, appliance, and computer	0.1020	0.0861	0.0462	0.0797	0.1267
Restaurants	0.0579	0.0529	0.0291	0.0477	0.0733
Miscellaneous	0.0966	0.0823	0.0464	0.0779	0.1242

TABLE 8. Annual Intangible Investments, Intangible Assets, and Profit Margins

This table presents the time series of the total number, sales revenue, operating expenses, intangible investments, recorded and unrecorded intangibles, and intangible adjusted and unadjusted profit margins of retailers.

Panel A: Intangible investments in comparison with sales, operating expenses, and profits (Unit: $ million)

Year	Number of firms	SALE	COGS	XSGA	XAD	XRD	InvOExp	IB
1995	144	411,218	293,058	85,807	5,246	3	14,584	10,297
1996	307	502,629	362,545	103,483	5,267	4	13,714	11,379
1997	321	633,594	458,093	128,264	7,869	38	10,397	13,989
1998	299	658,435	471,057	134,303	10,162	78	6,083	18,831
1999	281	770,357	554,603	156,119	10,139	291	10,591	19,611
2000	274	812,492	584,583	166,560	12,211	160	7,190	17,957
2001	303	972,139	689,451	209,208	16,258	146	16,887	21,307
2002	347	1,225,150	874,851	251,031	19,540	372	15,165	37,606
2003	326	1,305,993	935,328	264,082	21,314	422	18,393	43,013
2004	311	1,424,853	1,025,518	282,410	24,276	406	12,353	47,414
2005	308	1,530,401	1,103,460	302,546	25,205	589	14,148	55,112
2006	304	1,638,564	1,183,207	320,497	26,689	829	7,112	59,234
2007	281	1,694,241	1,224,050	334,022	26,507	1,062	13,400	55,238
2008	277	1,744,789	1,268,356	344,770	26,614	1,388	7,723	31,271
2009	277	1,782,783	1,284,614	355,904	25,325	1,615	23,130	51,175
2010	277	1,900,512	1,368,997	372,203	27,663	2,161	15,700	63,057
2011	261	1,969,743	1,432,818	375,757	26,660	3,168	22,426	61,235
2012	264	2,057,959	1,499,111	384,314	27,873	4,840	19,678	70,577
2013	268	2,178,265	1,590,122	407,044	29,906	6,896	18,964	71,086
2014	266	2,250,500	1,642,543	414,026	30,579	9,636	29,555	77,619
2015	260	2,391,301	1,732,906	448,618	31,739	12,925	18,672	81,425
2016	248	2,470,553	1,782,548	471,843	33,877	16,545	29,227	89,178
2017	235	2,521,950	1,813,768	490,920	37,032	23,125	22,238	92,539

Panel B: Recorded and unrecorded intangible assets (Unit: $ million)

Year	Assets recorded on the balance sheet			Unrecorded Intangible Assets			
	AT	GDWL	ROINTAN	Know	Ad	Org	UI
1995	230,601	4,716	2,352	3	5,246	14,584	19,833
1996	244,704	9,702	3,069	5	7,289	22,143	29,438
1997	311,050	14,987	4,156	41	12,398	26,790	39,229
1998	331,405	18,904	6,489	102	17,705	24,987	42,796
1999	379,115	20,761	18,662	350	18,514	28,806	47,671
2000	387,931	12,559	16,675	221	25,039	27,335	52,595
2001	484,193	17,532	19,813	272	33,715	39,486	73,473
2002	632,786	37,501	21,482	667	43,281	50,097	94,045
2003	661,832	49,826	14,179	866	49,430	56,463	106,759
2004	722,306	62,593	17,341	896	57,145	57,017	115,059
2005	781,634	79,629	19,964	1,175	58,361	57,196	116,732
2006	817,551	89,829	18,061	1,614	64,437	51,076	117,128
2007	899,143	114,455	35,812	2,137	68,258	52,966	123,362
2008	880,043	104,733	35,848	2,818	71,269	49,499	123,586
2009	926,123	108,251	37,583	3,502	72,560	62,435	138,497
2010	981,430	110,860	38,377	4,507	77,738	65,815	148,059
2011	986,491	118,619	40,156	5,924	74,607	73,386	153,917
2012	1,081163	149,198	58,938	8,809	76,589	72,640	158,039
2013	1,133,264	149,628	58,395	12,779	82,540	78,833	174,152
2014	1,207,425	175,027	82,467	18,198	83,634	90,773	192,604
2015	1,311,979	204,135	102,598	25,105	86,985	89,722	201,812
2016	1,368,006	220,190	110,309	33,365	91,538	99,401	224,304
2017	1,416,469	238,347	120,383	45,475	95.330	98,657	239,462

Panel C: Profit margin with and without unrecorded intangible assets (Unit: $ million)

Year	Intangible Investment (iINV)	Intangible-adjusted IB (iIB)	IB/SALE	iIB/SALE	Difference
1996	18,985	25,715	2.26%	5.12%	2.85%
1997	18,304	25,457	2.21%	4.02%	1.81%
1998	16,323	25,691	2.86%	3.90%	1.04%
1999	21,021	29,758	2.55%	3.86%	1.32%
2000	19,561	25,532	2.21%	3.14%	0.93%
2001	33,291	40,795	2.19%	4.20%	2.00%
2002	35,077	53,570	3.07%	4.37%	1.30%
2003	40,129	58,620	3.29%	4.49%	1.20%
2004	37,035	56,559	3.33%	3.97%	0.64%
2005	39,942	64,497	3.60%	4.21%	0.61%
2006	34,630	62,778	3.61%	3.83%	0.22%
2007	40,969	64,195	3.26%	3.79%	0.53%
2008	35,725	33,172	1.79%	1.90%	0.11%
2009	50,070	66,896	2.87%	3.75%	0.88%
2010	45,524	70,994	3.32%	3.74%	0.42%
2011	52,254	73,185	3.11%	3.72%	0.61%
2012	52,391	81,716	3.43%	3.97%	0.54%
2013	55,766	84,143	3.26%	3.86%	0.60%
2014	69,770	100,167	3.45%	4.45%	1.00%
2015	63,336	93,002	3.41%	3.89%	0.48%
2016	79,649	113,893	3.61%	4.61%	1.00%
2017	82,395	113,836	3.67%	4.51%	0.84%

TABLE 9. Regressions of Operating Expenses on Sales and Intangible Investment Texts

This table presents the parameter estimates from cross-sectional regressions to separate intangible investments from operating expenses using financial statement texts. The numbers in parentheses are *t*-statistics, and ***, **, and * mean the statistical significance at the one, five, and ten percent levels, respectively.

Panel A: Using both XSGA and COGS

Dependent variable: OExp/AT	SALE/AT	D_SDrop	D_Loss	$(IDIS_t - IDIS_{t-1})*I^P$	Adj-R^2(%)
All retailers with subsector and year dummies	1.57 (363.78)***	0.02 (0.42)	0.19 (4.83)***	0.21 (2.81)***	95.38
Subsectors (two-digit SIC) with year dummies					
Hardware (52)	0.97 (184.49)***	0.06 (2.66)***	0.27 (11.75)***	0.08 (1.66)*	99.46
Department stores (53)	1.04 (177.04)***	0.05 (3.80)***	0.14 (9.46)***	-0.01 (-0.31)	98.32
Grocery (54)	0.99 (293.13)***	0.00 (0.45)	0.09 (11.93)***	0.00 (0.19)	99.46
Auto and gas (55)	1.00 (218.15)***	0.02 (1.35)	0.14 (8.59)***	0.01 (0.22)	99.09
Clothing (56)	1.02 (117.36)***	0.07 (5.23)***	0.20 (14.03)***	0.08 (3.15)***	94.54
Furniture/computer (57)	1.05 (97.16)***	0.04 (1.70)*	0.23 (9.46)***	0.07 (1.99)**	96.01
Restaurants (58)	1.58 (639.11)***	0.01 (0.38)	0.10 (2.62)***	0.35 (3.53)***	99.62
Miscellaneous (59)	1.64 (89.27)***	0.06 (0.47)	0.38 (3.13)***	0.27 (1.32)	82.38

Panel B: Using XSGA only

Dependent variable: XSGA/AT	SALE/AT	D_SDrop	D_Loss	$(IDIS_t - IDIS_{t-1})*I^P$	Adj-R^2(%)
All retailers with subsector and year dummies	0.71 (235.71)***	0.04 (1.48)	0.08 (3.07)***	0.05 (1.06)	89.74
Subsectors (two-digit SIC) with year dummies					
Hardware (52)	0.19 (19.46)***	0.02 (0.57)	0.25 (5.93)***	0.21 (2.25)**	70.62
Department stores (53)	0.24 (20.35)***	0.08 (3.16)***	0.14 (4.56)***	0.06 (1.03)	46.32
Grocery (54)	0.23 (24.76)***	0.02 (0.84)	0.03 (1.24)	-0.05 (-0.95)	57.86
Auto and gas (55)	0.04 (4.27)***	0.02 (0.80)	0.16 (5.37)***	0.00 (0.04)	9.77
Clothing (56)	0.27 (25.22)***	0.04 (2.68)***	0.13 (7.58)***	0.06 (1.94)*	48.31
Furniture/computer (57)	0.19 (14.68)***	0.06 (2.22)**	0.17 (5.72)***	-0.12 (-2.63)***	43.15
Restaurants (58)	0.78 (276.89)***	0.01 (0.15)	-0.01 (-0.32)	0.33 (2.91)***	98.05
Miscellaneous (59)	0.35 (38.85)***	0.14 (2.87)***	0.31 (6.59)***	-0.03 (-0.38)	48.87

TABLE 10. Textual Analysis of Financial Firms

This table presents textual analysis of financial firms' 10Ks matched with numerical information in Compustat and XBRL tags. There are 11,746 firm-years for the fiscal years 2009 – 2021. Panel A presents summary statistics and Panel B shows regressions to test the relation between intellectual property dictionary (iPDic) counts and operating expenses on communication, information technology, and data scaled by sales. The numbers in parentheses are *t*-statistics, and *** means the statistical significance at the one percent level.

Panel A: Summary statistics

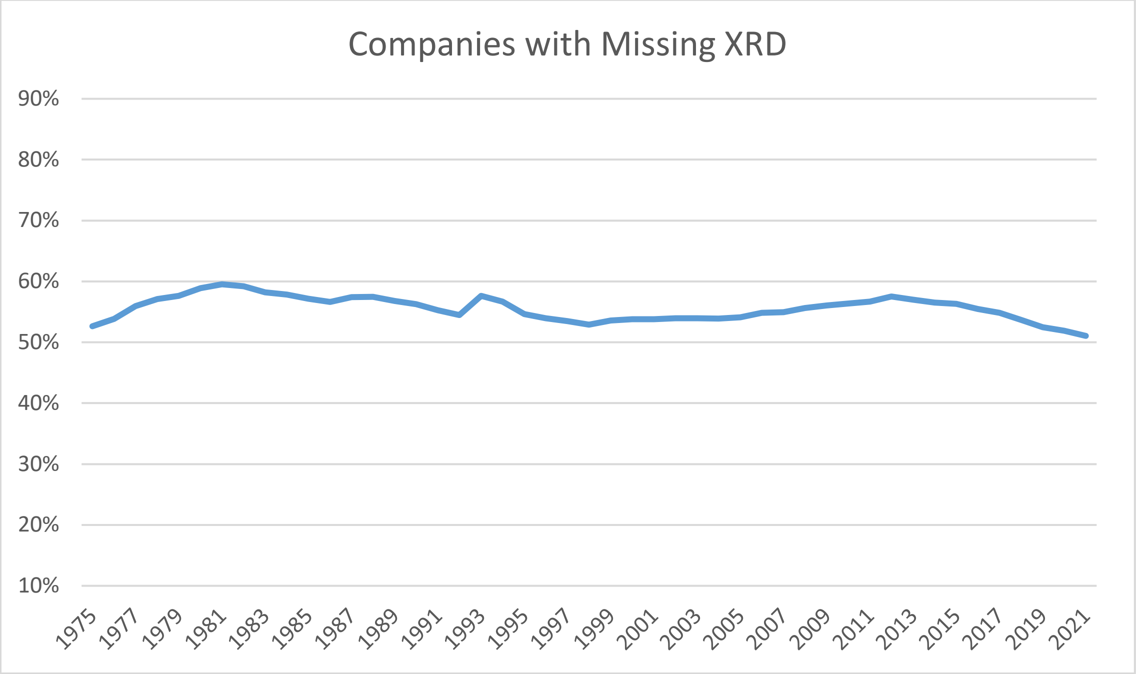
Variable	Number	Mean	Standard Deviation	Percentiles		
				25th	Median	75th
XBRL tags scaled by SALE (%)						
COM	671	1.60	1.31	0.88	1.23	1.81
CIT	841	4.13	2.37	2.43	3.86	5.32
ITDP	2,376	3.61	1.96	2.23	3.17	4.65
COMITDP	3,888	3.38	2.13	1.78	2.97	4.51
XCOM scaled by SALE (%)	2,923	2.04	2.40	0.98	1.51	2.33
Number of words (NW)	11,746	66,884	41,952	45,307	57,539	75,223
Intellectual property dictionary (iPDic) counts	11,636	42	59	15	26	45
iPDic/NW (%)	11,636	0.06	0.09	0.03	0.04	0.07

Panel B: Regressions of $\text{iPDic}_t/\text{NW}_t$ (%) on previous expenses

Previous three years expenses scaled by the total SALE for the same period	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	0.07 (2.72)***	0.09 (15.36)***	0.15 (7.71)***	0.12 (5.80)***	0.09 (26.75)***
COM	-0.66 (-3.61)***				
CIT		0.27 (2.63)***			
ITDP			-0.21 (-4.10)***		
COMITDP				-0.07 (-1.77)*	
XCOM					0.09 (1.55)
Depository institutions (2-digit SIC:60)	-0.01 (-0.43)	-0.04 (-9.08)***	-0.09 (-4.71)***	-0.06 (-3.05)***	-0.03 (-12.27)***
Non-depository credit institutions (61)		-0.02 (-2.25)**	0.02 (0.78)	-0.03 (-1.20)	
Brokers and dealers (62)	0.07 (2.46)**		-0.02 (-1.22)	-0.01 (-0.44)	
Adj-R^2(%)	57.78	16.16	21.69	20.11	12.15

Figure 1. R&D Expenditures in Compustat

This figure shows i) the proportion of firms that do not have R&D expenditures (XRD) in Compustat by fiscal year for the 403,510 firm-years with non-missing sales revenue (SALE) from 1975 to 2021, and ii) total XRD scaled by total SALE in fiscal year 2021 by industry based on SIC codes using Fama-French 12 industry classification.



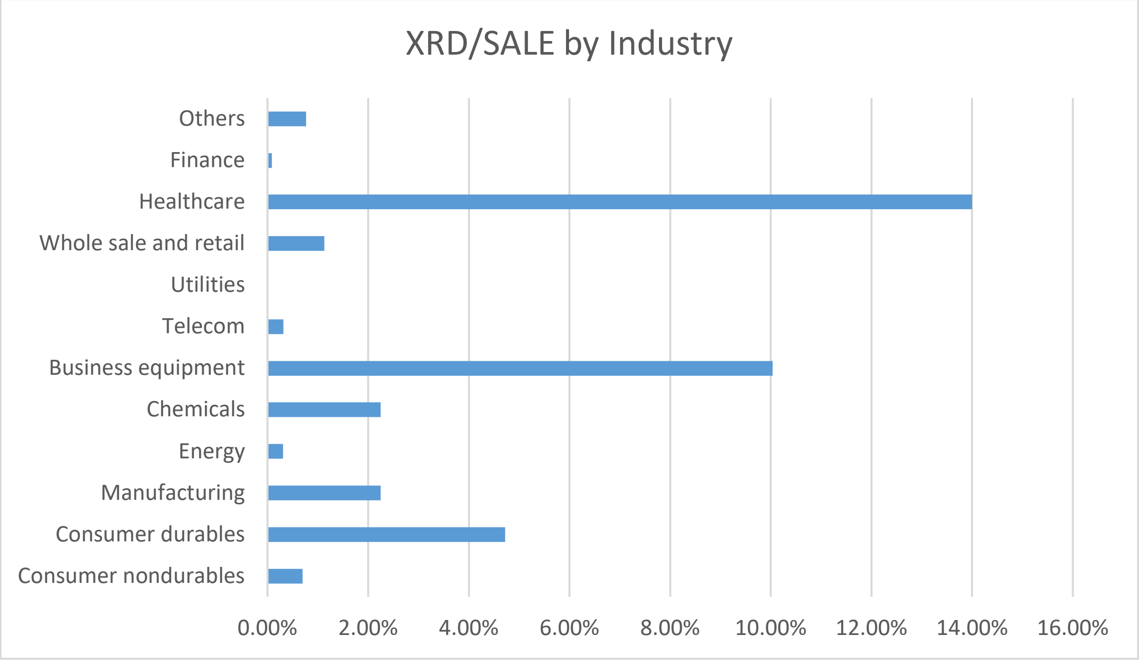


Figure 2. The Growth of Patents Issued to Firms in the Finance Industry

This figure presents the number, market values, and the received citations of patents issued to firms in the finance industry in proportions of those in all CRSP firms each year. The patent values are estimates of Kogan et al. (2017).

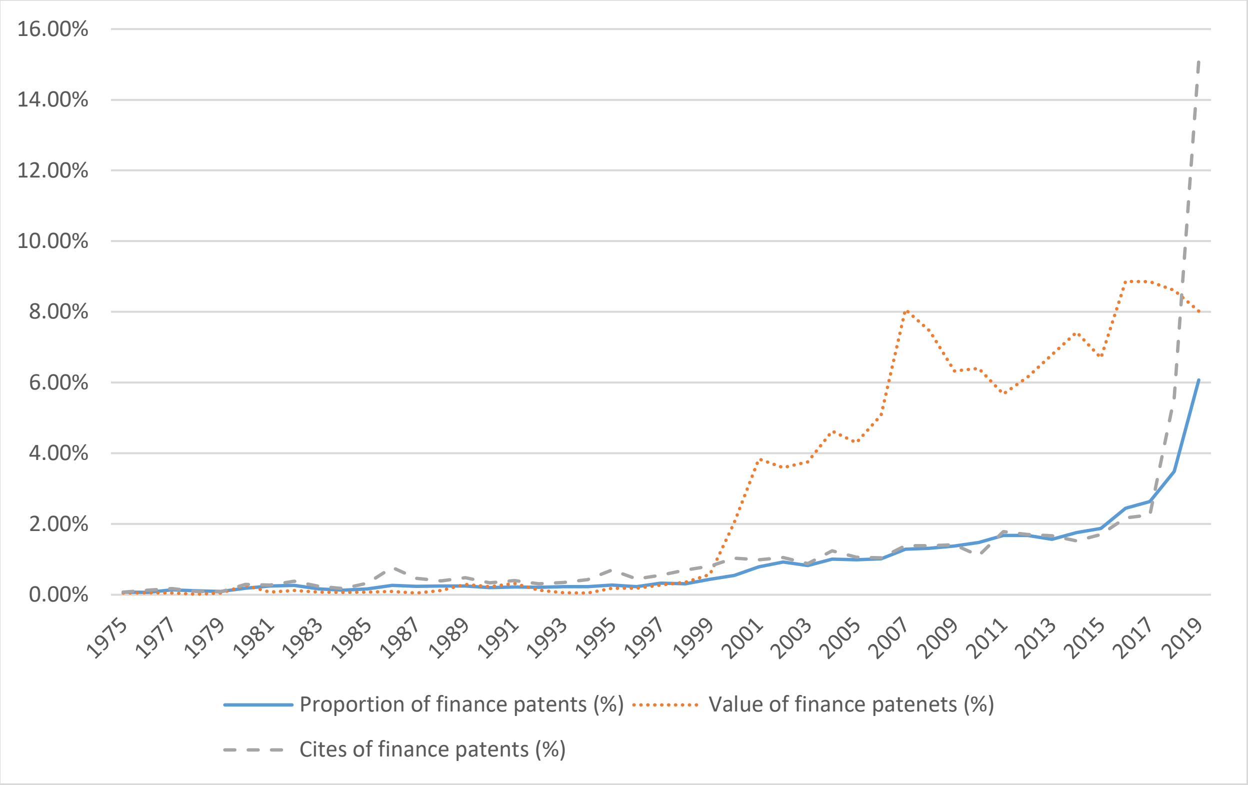
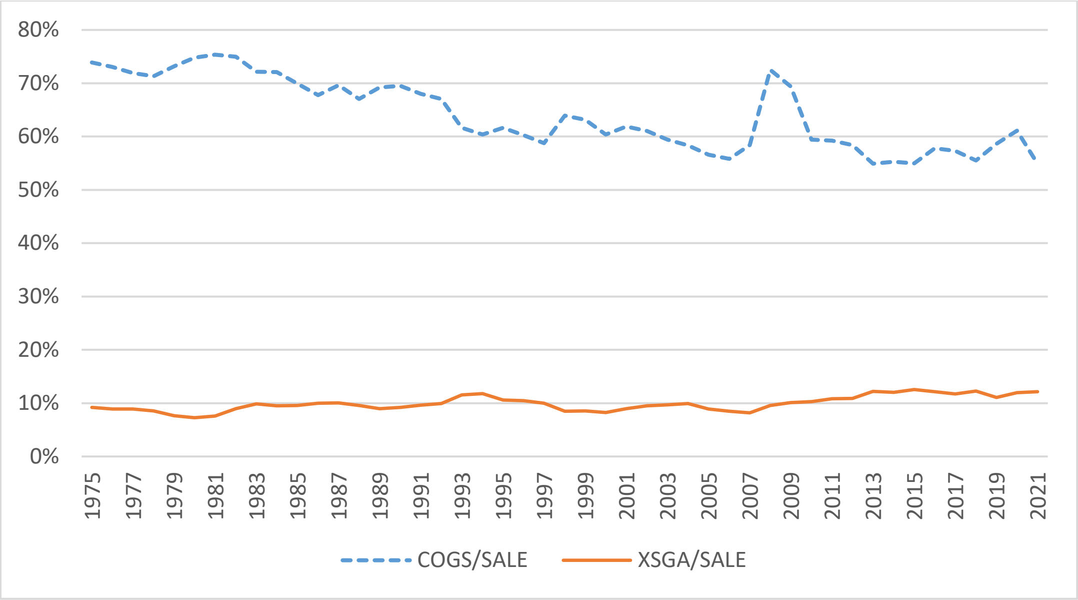


Figure 3. The operating expenses data in Compustat for the Finance Industry: 1975 - 2021

This figure presents (a) the time series of cost of goods sold (COGS), selling, general and administrative expenses (XSGA), scaled by sales revenue (SALE) and (b) the proportion of firms reporting communications expense (XCOM) in the Compustat database for the firms in the finance industry (SIC code: 6000 – 6999).



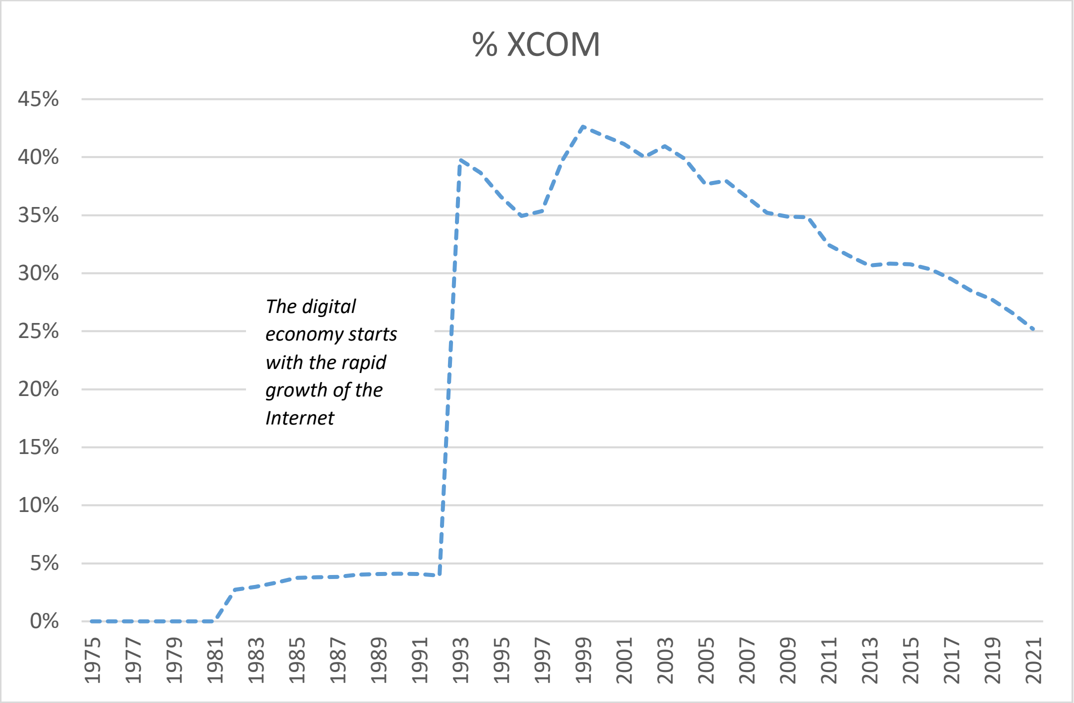


Figure 4. Intangibles Dictionary to Estimate the Unrecorded Intangible Assets of Retailers

I developed this word cloud using 10-Ks of retailers investing heavily in e-commerce. I sorted out the words and phrases used to describe intangible investments and define the list as the Intangibles Dictionary (iDic). I search the entire retail 10-Ks for the words and phrases in iDic using Python and form this word cloud that visualizes the intangibles word frequencies in the retail annual reports filed with SEC EDGAR during 1994-2018.

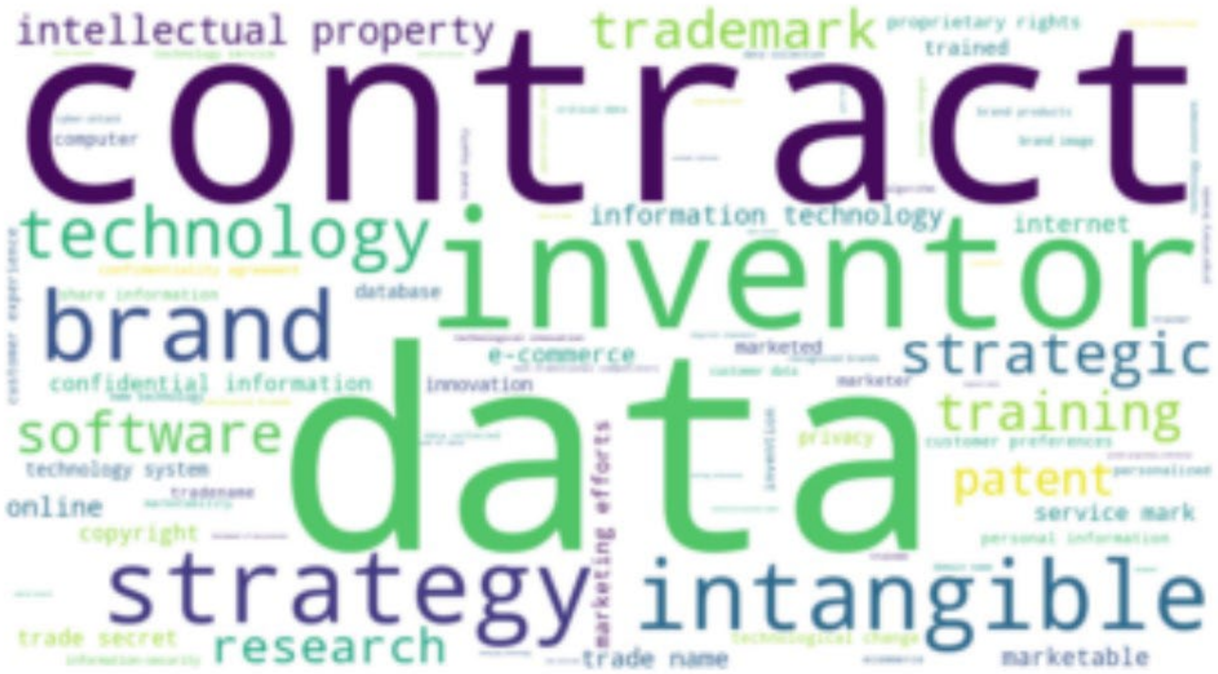
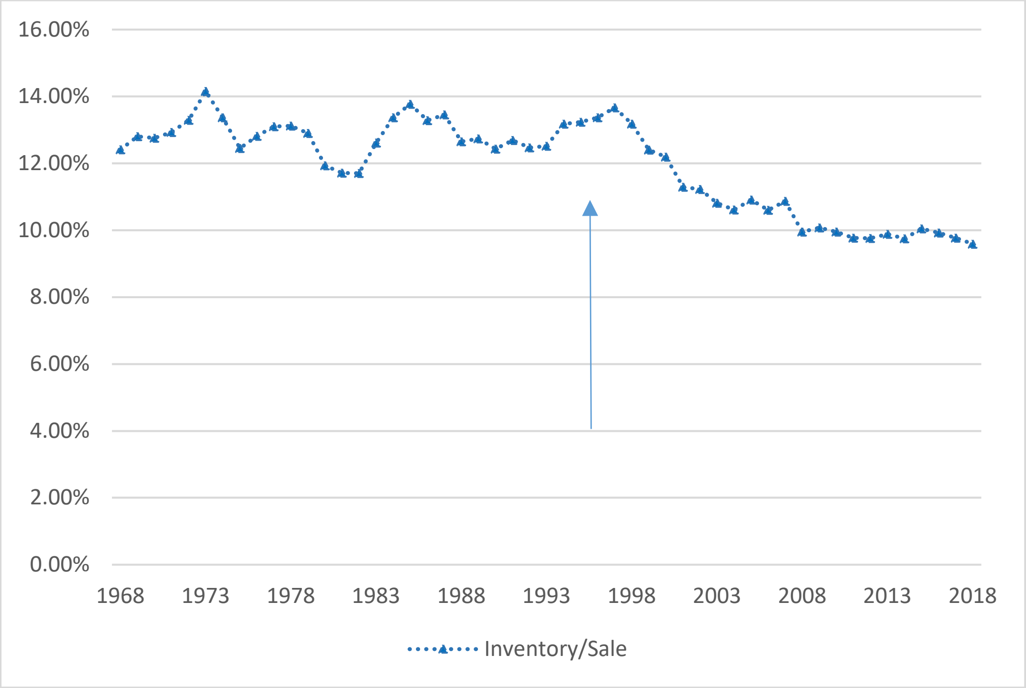


Figure 5. Innovation of Retailers the digital economy

Retailers show significant improvements in efficiencies and productivity measured by Inventory/Sale and Sale/employee in the digital economy that started in the mid-1990s.

(a) Inventory management



(b) Sales revenue per employee

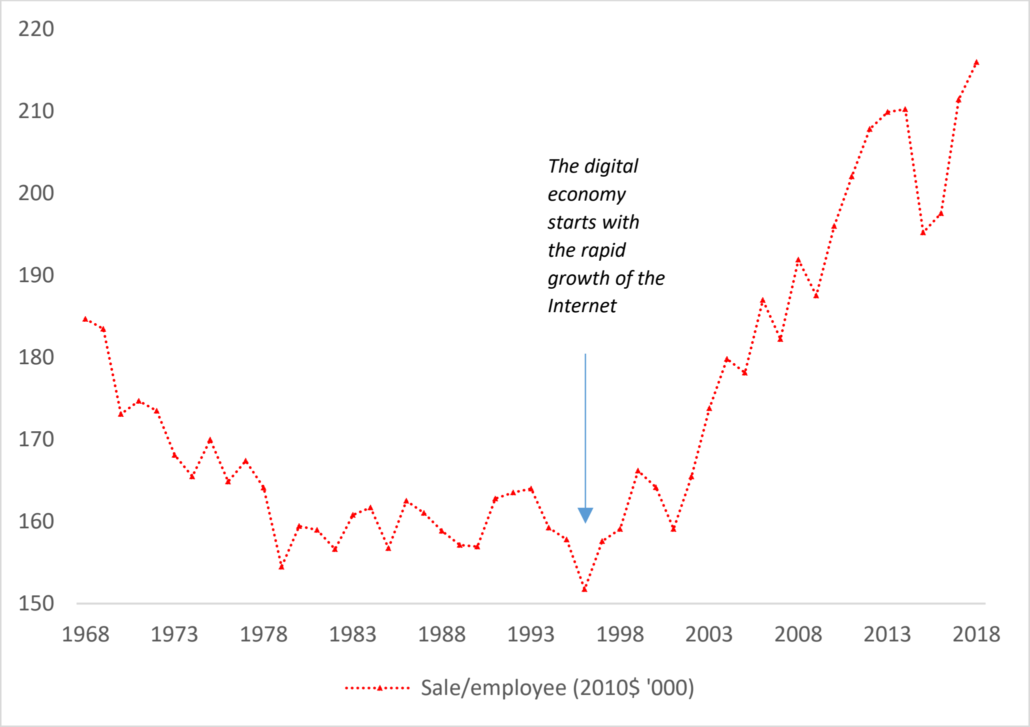
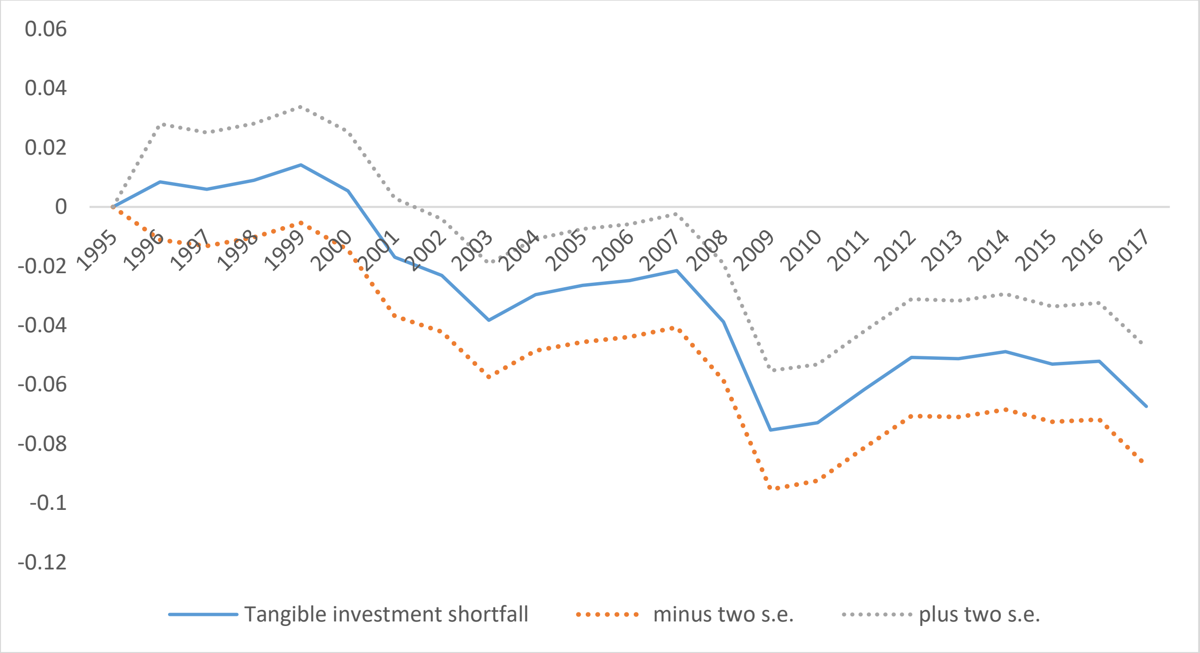


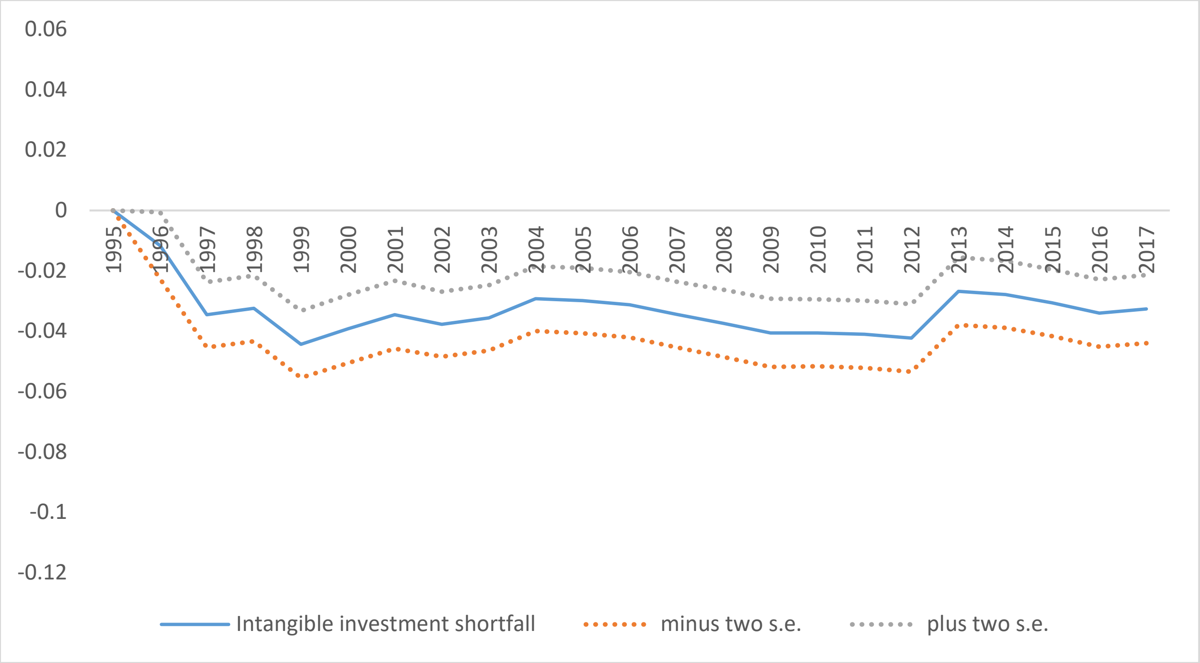
Figure 6. Comparing Tangible and Intangible Investment Shortfalls of Retailers

This figure compares tangible investment shortfall with intangible and intangible-adjusted tangible investment short fall. Tangible investment shortfall in (a) is the time effects in the regression of CAPX/PPEGT on Tobin's Q and EBITDA/PPEGT. Intangible investment shortfall in (b) is the time effects in the regression of iInv/iAT on Total Q and iEBITDA/iAT. Intangible-adjusted tangible investment shortfall in (c) is the time effects in the regression of CAPX/iAT on Total Q and iEBITDA/iAT.

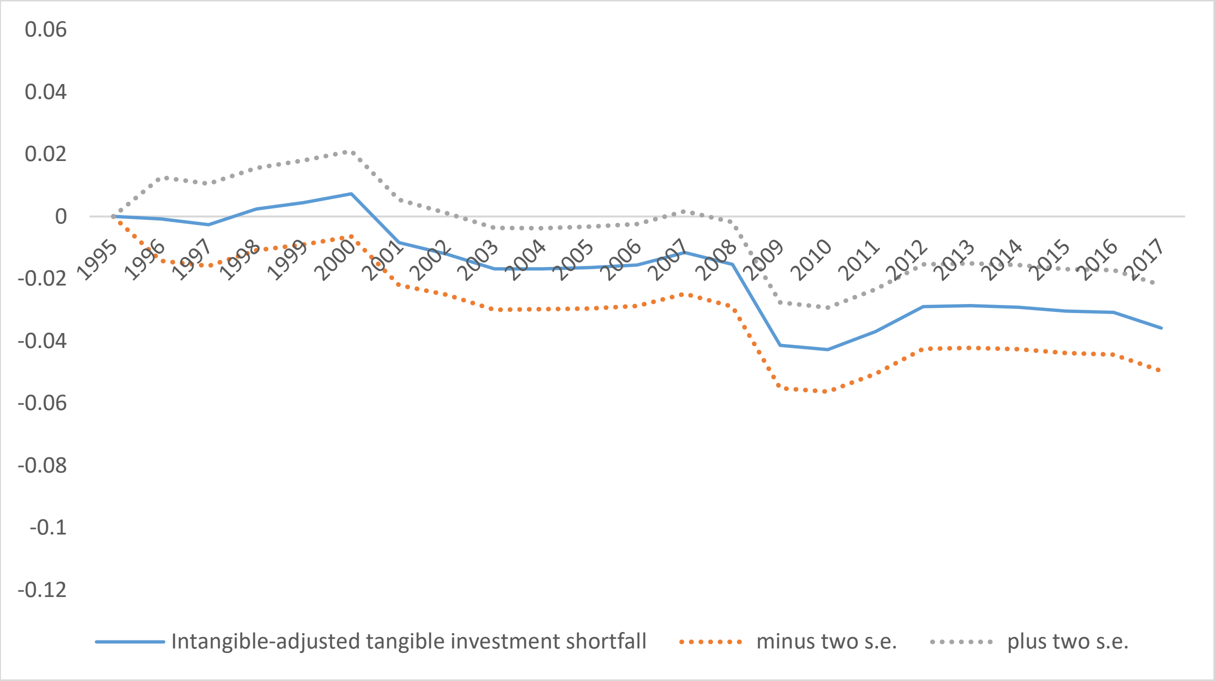
(a) Tangible investment shortfall



(b) Intangible investment shortfall



(c) Intangible-adjusted tangible investment shortfall


Intangible-adjusted tangible investment shortfall
minus two s.e.
plus two s.e.