New Directions in Empirical Asset Pricing: Information, Innovation, and Stock Returns

by

Julian Vahl

BSc, University of Tuebingen, 2013

MSc, Vienna University of Economics and Business, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Business Administration - Finance)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

New Directions in Empirical Asset Pricing:
Information, Innovation, and Stock Returns

submitted by **Julian Vahl** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Business Administration - Finance**.

Examining Committee:

Murray Carlson, Professor, Business Administration - Finance, UBC
Co-Supervisor

Adlai Fisher, Professor, Business Administration - Finance, UBC
Co-Supervisor

Sandra Chamberlain, Associate Professor, Business Administration - Accounting, UBC
University Examiner

Kevin Song, Professor, Economics, UBC
University Examiner

Alberto Teguia, Assistant Professor, Business Administration - Finance, UBC
Supervisory Committee Member

Additional Supervisory Committee Members:

Werner Antweiler, Associate Professor, Business Administration - Strategy and Business Economics, UBC
Supervisory Committee Member

Abstract

This dissertation is a collection of three essays that explore new directions in empirical asset pricing.

The first essay studies the role of textual information in analyst reports. I show that analysts use the report text to convey soft information that has not yet been incorporated into their numerical forecasts. A simple tone measure predicts forecast revisions and forecast errors several periods ahead. Market prices quickly and adequately absorb the soft earnings information in analyst tone after the report publication. I demonstrate that analyst tone can be used to measure the saliency of upside or downside risks.

The second essay proposes a novel deep learning approach to identify predictors of announcement returns from text. The method reveals that stock returns on earnings announcement days are predictable using analyst reports published weeks before the announcements. The identified predictors perform well for several years out-of-sample but eventually vanish. The predictability arises from a persistent underreaction to firm-specific news. A portfolio strategy based on out-of-sample announcement predictions earns large significant alpha. The findings are consistent with biased expectations and not in line with common risk-based explanations.

The third essay studies the pricing of technological innovators in the stock market. It shows that technological innovators are priced differently, earning high stock returns controlling for standard factors, with less punishment for high capital investment and weak profitability. We create the persistent new firm variable patent intensity (PI), patents received divided by market capitalization, available from 1926. Aged PI portfolios and standard factors show high alpha and low profitability lasting more than a decade past formation for firms with high patenting intensity. Adding an expected growth factor, alphas become insignificant at most horizons, and loadings show large but declining growth, aggressive and increasing investment, and weak but improving profitability. The essay discusses partly unifying interpretations of some important factor models and the essential role of expected growth.

Lay Summary

This thesis is a collection of three essays that studies the relationship between information, innovation, and stock returns. The first essay studies how stock market analysts use language to transmit information about firms' future earnings to market participants. The second essay develops a novel machine learning method to extract information about stock returns from text documents and investigates how markets learn to incorporate this information into stock prices. The last essay demonstrates that the stock prices of innovative firms behave differently from the prices of non-innovative firms.

Preface

Chapters two and three of this dissertation are my original independent work. Chapter four is joint work with several colleagues from my time at UBC. My advisor Adlai J. Fisher, my former Ph.D. colleague Jiri Knesl, and I have contributed equally to the writing and empirical analysis in the manuscript that forms this chapter.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Market participants continuously collect, process, and evaluate any available information to form expectations about future earnings and returns of listed firms. Stock prices summarize the market's information about future earnings and returns and constantly move up and down to adjust to changes in these expectations. Collecting and processing information is costly, and forming the right expectations about future earnings based on present information is a complex task. Since the onset of financial research, scholars have studied whether markets perform this task efficiently and immediately. My thesis is a collection of three essays that each shed new light on the way markets process information about future earnings and returns.

In the first essay, I study how stock market analysts, an important group of information providers for institutional investors, transmit information about future earnings to market participants through text. While analysts also provide numerical estimates about future earnings, I demonstrate that the tone of the written reports that accompany the numerical estimates contains earnings information above and beyond the numerical estimates. Analyst tone predicts forecast errors and forecast revisions several quarters ahead, suggesting that analysts use language to transmit hard-to-quantify information. I show that markets react promptly and adequately to the publication of reports with tonal information. Furthermore, I suggest that analyst tone can be used to measure the salience of positive versus negative events, which can be used to predict over- and underreaction patterns to future news.

Motivated by the findings of the first essay, the second essay develops a new methodology to extract pricing-relevant information from text. I present a deep learning architecture that is able to extract and quantify mispriced information from any set of text documents associated with an individual firm. Using the same dataset of analyst reports as in the first essay, I show that markets underreact to particular types of textual information. The predictable underreaction can be used to form profitable trading strategies on earnings announcement days several weeks after the publication of the reports. Continuous re-training of the deep learning model reveals that markets eventually learn to incorporate the particular type of information that has led to mispricing in the past. However, they only do so slowly over the course of several years.

In the third and last essay, I study a particular set of corporate information and its role in determining stock prices: firms' innovation activities. I use firms' patent grants and market prices to construct a measure of patenting intensity that allows me to study the prices of innovating firms over a time span of almost 100 years. Technological innovators appear to be priced differently in the stock market. Firms with high patenting intensity earn high abnormal returns relative to their no-innovation or low-innovation counterparts. The abnormal return can be explained by an apparent undervaluation of firms with high patenting activity, suggesting that markets do not adequately incorporate the information contained in firms' patenting activity into prices. Furthermore, I show that while innovating firms accounted for more than half of the total market capitalization of the U.S. stock market for most of the past century, several popular predictors of the cross-section of returns do not work in the sub-sample of innovating firms. Together, the findings in this essay suggest that innovating firms are priced differently in the market.

Although the topic of information and stock returns is present in all three essays, each essay investigates a different research question. Therefore, chapters were designed to be self-contained. I leave a more exhaustive discussion of the research question, methodology, and contribution to the introduction specific to each chapter.

Chapter 2

Beyond the Numbers: Earnings Information in Analyst Tone

2.1 Introduction

Equity research analysts typically publish earnings estimates and stock recommendations in the form of an analyst report. In addition to the quantitative estimates such as earnings forecasts and target prices, these reports contain a significant amount of text.[1] While an extensive literature examines the information content in analyst forecasts and other quantitative estimates and opinions expressed in the analyst reports, little is known about the value of the textual information that typically accompanies these estimates. Despite the fact that the quantitative estimates are often available online or through subscription services such as the Bloomberg Terminal, asset managers often pay high five-digit fees for the access to the equity research platform of a single brokerage house[2]. This suggests that the value of analyst reports goes well beyond the quantitative information therein.

This paper explores whether analyst tone contains information about future earnings and affects information processing in financial markets in two parts. In the first part of the paper, I show that a simple measure of analyst tone - a dictionary-based sentiment measure - captures earnings information in analyst reports. Here, I focus on the information content in the report text and its incremental value for predicting future earnings over the numerical analyst forecasts. In the second part of the paper, I conjecture that analyst tone does not only represent soft earnings information, but also the analyst's attention to particular news, and investigate whether analyst attention impacts

[1] The median analyst report in the sample contains 682 words of text, excluding tables, figures, imprints, disclaimers, etc. See Section 2.2.3 for more details.

[2] https://www.bloomberg.com/professional/blog/put-price-investment-research-2/

the price reaction to future news in financial markets.

I show that analyst tone conveys valuable information about future earnings. Positive analyst tone predicts positive earnings surprises relative to the numerical estimates in the same analyst reports, and negative analyst tone predicts negative earnings surprises. The predictability is particularly strong for further-ahead periods. My findings are consistent with analysts using the report text to transmit soft information that is difficult or costly to quantify at the time of the publication. Furthermore, analyst tone predicts revisions of further-ahead forecasts after the earnings announcement. This suggests that the soft information in the reports will be transformed into hard information once additional information arrives.

Stock markets react positively to reports with positive tone and negatively to reports with negative tone[3]. Pre-publication returns several days before the publication reveal that analysts use the report text to respond to recent events that are already accounted for in the stock prices. Approximately half of the trailing ten-day cumulative abnormal return on the report publication day is realized before the publication day itself, suggesting that analyst tone reflects previously publicized information and new information in roughly equal parts. The lack of a post-publication drift in the days after the publication suggests that the market quickly incorporates the tonal information into stock prices. Despite the ability to predict earnings surprises, analyst tone does not predict earnings announcement reactions, suggesting that markets efficiently incorporate the earnings information in analyst tone into prices.

In the second part of the paper, I study whether analysts' attention to positive versus negative events in their reports directs investor attention in financial markets. Assuming that each analyst report is not an exhaustive summary of the analysts' private information, but a subset of the information available to the analyst that the analyst decided to present to the readers, analyst tone reveals the analyst's choice to direct their own or their readers' attention to certain soft information.

Analyst tone is low when analysts primarily discuss negative scenarios or events in their report, and high when they focus on positive events. I argue that analysts' focus on positive vs negative scenarios is a useful proxy of directional investor attention. First, the target audience for analyst reports are institutional investors. To the extent that institutional investors pay attention to analyst publications, events and scenarios discussed in analyst reports are likely to be salient for a group of investors that is responsible for the majority of the daily trading volume. Second, analysts are skilled investment professionals themselves that engage in frequent exchange with institutional investors, meaning that analyst attention is more likely to be in line with other investment professionals' attention. Third, the objective of analysts is to provide investors with value-relevant information. As such, the decision to publish a report about a certain firm and topic is not random: the information

[3]A similar result has been shown by Huang et al. (2014) in a smaller sample and less general setup.

displayed in the reports is likely to be the information that analysts think investors should pay attention to. This paper remains agnostic about the drivers of (in)attention and focuses on the effect of varying investor attention for information processing in financial markets.

I find that pre-announcement analyst tone predicts post-earnings announcement drift and reversal conditional on the sign of the earnings news. If the news is in line with the pre-announcement analyst tone, e.g. a negative surprise following a negative consensus tone, stock prices tend to overreact to the news. Conversely, if the news is not in line with the pre-announcement analyst tone, e.g. a positive surprise following a negative consensus tone, stock prices tend to underreact. After the initial announcement reaction, prices show a strong drift or reversal pattern over the next 10-15 trading days, which in some cases persists for up to 90 days. This suggests that stock markets overreact to salient events and underreact to non-salient events. The fact that pre-announcement analyst tone in itself does not predict whether the stock price will drift or reverse after the announcement rules out explanations related to symmetric informational frictions such as slow information diffusion or sticky expectations, especially those frictions that can be linked to rather constant firm characteristics (e.g. firm size). Instead, the findings highlight that announcement drifts are at least partially determined by an interaction between the realized surprise and the directional attention of investors before the announcement.

The observed predictable drift and reversal patterns can be exploited by a simple portfolio trading strategy. Since low analyst tone firms underreact to good news and overreact to bad news, abnormal returns are always positive after the initial announcement reaction for negative analyst tone firms, irrespective of the sign of the announcement reaction. Similarly, abnormal returns are always negative following the initial announcement reaction for positive analyst tone firms. A trading strategy that buys low analyst tone stocks and sells high analyst tone stocks after an earnings announcement earns a Fama-French four-factor alpha of 38bps per week. The profits are primarily generated by the long side of the portfolio, which suggests that such a trading strategy would be easy to implement and the trading profits are not an artifact of short-selling constraints.

To the best of my knowledge, I collect by far the largest sample of analyst reports that has been used in the financial literature up to today[4]. I am able to match the text corpus of 1.6 million analyst reports for S&P1500 firms published between 1982 and 2019 with quantitative earnings forecasts from I/B/E/S. The vast sample allows me to conduct a more extensive analysis of the textual information in these reports than in the previous literature. The literature on analyst reports is still small, and only recently gained some traction due to the widespread availability of cheap

[4]Huang et al. (2014) use a sample of 363,952 reports for S&P500 firms, Huang et al. (2017) study 159,210 reports for S&P500 firms, De Franco et al. (2015) study 356,463 reports, Hsieh et al. (2016) study 2,164 reports, and Gultekin et al. (2019) study 724,829 reports.

processing power and general interest in textual information in finance and economics.[5]

My paper contributes to several strands of the literature on analysts, soft information, and attention in financial markets. First, I contribute to the literature on soft information in analyst reports. The information content in analyst reports has first been studied by Huang et al. (2014), who document that analyst reports with positive sentiment trigger positive stock market reactions and conclude that textual sentiment in analyst reports carries value-relevant information for stock prices. Huang et al. (2017) attest analysts an important role in interpreting corporate disclosures.

Second, I provide further evidence on the interaction between earnings estimates and soft information. Previous studies have shown that the textual sentiment of third-party text corpora such as newspapers (e.g. Tetlock, 2007; Tetlock et al., 2008) or conference calls (Price et al., 2012) predicts forecast errors with a positive sign. This has often been interpreted as evidence of the lack of attention of analysts to new information or irrationally sticky forecast estimates (Bouchaud et al., 2019). By showing a similar relationship between forecast errors and the sentiment of the forecasters themselves, I argue that it is plausible that analysts rationally decide to exclude certain information from their estimates, and choose to transmit hard-to-quantify information via the report text. An overview of the literature on soft information in financial markets can be found in Liberti and Petersen (2019)

Third, I contribute to the literature on limited attention of market participants. Dellavigna and Pollet (2009) and Hirshleifer et al. (2009) suggest that post-earnings announcement drifts can be attributed to limited investor attention. Directional attention of investors has been studied in an experimental setting by Kuhnen (2015), who shows that subjects overreact to negative news when they are in a loss domain. Subjective attention and its impact on information processing have a long history in the psychology literature. Tversky and Kahneman (1973) show that subjective probabilities are influenced by the ease with which certain states come to mind. Tversky and Kahneman (1974) suggest that the subjective probability of an event that is easily retrievable is likely to be overstated.

Fourth, I add to the broader literature on analyst forecasts and analysts' reaction to news. Existing literature has largely focused on numerical forecasts provided by analysts. Among others, Bondt and Thaler (1990) show that analysts overreact to certain types of information while Abarbanell and Bernard (1992) argue that analysts underreact to other types of information. Bradshaw (2011) provides a survey of the literature on numerical analyst forecasts. Suggesting that numerical forecasts are only an incomplete measure of analyst expectations, my paper sheds new light on many of these findings.

The rest of the paper is organized as follows. Section 2.2 introduces the data and defines the measure of analyst tone and expectations. Section 2.3 discusses properties of analyst tone. Sec-

[5]Loughran and Mcdonald (2016) and Gentzkow et al. (2019) provide comprehensive surveys of the use of textual data in finance, accounting, and economics beyond analyst reports.

tion 2.4 investigates the information content in analyst tone. Section 2.5 analyzes how analyst tone shapes the market's reaction to news. Section 3.8 concludes.

2.2 Data and definitions

2.2.1 Earnings forecasts

Analysts publish forecasts and reports throughout the year, but releases are typically clustered around scheduled earnings announcements. This makes earnings announcements the ideal setting for our study because of the wide cross-section of observable pre- and post-announcement expectations. I obtain analyst expectations of quarterly earnings per share from I/B/E/S. I measure an analyst j's prior expectations as the latest recorded forecast within 45 days prior to the earnings announcement date of firm i and time t, and the posterior expectation as the last recorded forecast within 45 days after the earnings announcement date. I use a 45-day window around the earnings announcement dates following Bouchaud et al. (2019), who document that 45 days is the median time across analysts to issue an earnings forecast after an earnings announcement in the I/B/E/S dataset. Consensus estimates are the average of the individual estimates and denoted by $\mathbb{E}_{i,t^-}$ and $\mathbb{E}_{i,t}$, where subscript t^- indicates prior estimates and t indicates posterior estimates.[6]

Using $C_i^{t+\tau}$ to denote the earnings per share of firm i in period $t+\tau$, let $N_{i,t}$ be the consensus forecast revision scaled by the stock price at the beginning of the 45-day window,

$$N_{i,t}^{t+\tau} = \frac{\mathbb{E}_t[C_i^{t+\tau}] - \mathbb{E}_{t^-}[C_i^{t+\tau}]}{P_{i,t^-}}. \tag{2.1}$$

I use FE_t to denote the forecast error relative to time t expectations. In particular, $FE_{i,t}^{t+\tau}$ is the period $t+\tau$ forecast error for firm i at time t,

$$FE_{i,t}^{t+\tau} = \frac{C_i^{t+\tau} - \mathbb{E}_t[C_i^{t+\tau}]}{P_{i,t^-}}. \tag{2.2}$$

Again, the subscript t here refers to expectations measured after the release of period t earnings, with a delay of up to 45 days. In contrast to the forecast *revision*, the forecast *error* compares the current earnings expectation with the future realized earnings. Therefore, it is forward-looking and

[6]Unreported robustness checks confirm similar results when using the median instead of the average of individual expectations.

can only be observed at $t+\tau$. I also use forecast errors relative to pre-announcement expectations,

$$FE_{i,t^-}^{t+\tau} = \frac{C_i^{t+\tau} - \mathbb{E}_{t^-}[C_i^{t+\tau}]}{P_{i,t^-}}. \tag{2.3}$$

Throughout the paper, all revision and error variables are standardized to mean zero and unit standard deviation. Summary statistics are shown in Table 2.1 Panel A.

2.2.2 Recommendations and target prices

I merge consensus forecast information with buy/sell recommendations and target prices from I/B/E/S. I convert buy/sell recommendations to a numerical scale by assigning the numbers 1-5 to the recommendations ”strong sell”, ”sell”, ”neutral/hold”, ”buy”, and ”strong buy”, respectively. The consensus recommendation is obtained by following the aggregation procedure for consensus expectations: I first measure analyst j's prior recommendation as the latest recorded recommendation of analyst j within 45 days prior to the earnings announcement date of firm i and time t. The consensus recommendation Rec_{i,t^-} is then given calculated as the average of the individual recommendations.

Target prices are expressed as returns to make them comparable across firms with different price levels. The target price implied return (TPIR) is the 12-months target price divided by the stock price at the time of announcement of the target price minus one. As with earnings expectations and recommendations, I measure analyst j's prior TPIR as the latest recorded TPIR within 45 days prior to the earnings announcement date of firm i and time t. Consensus target price implied returns $TPIR_{i,t^-}$ are the average of the individual TPIRs.

2.2.3 Analyst reports

Data collection and matching

I collect analyst reports for all historic S&P1500 constituents from Thomson One Investext from April 1982 to June 2019. The raw database contains 3.6 million reports. I remove all reports that do not contain firm-specific analyst opinions from the database. In particular, I remove reports from contributors that provide machine-generated reports, or reports that are not equity-focused opinion pieces (e.g. company descriptions, merger news, drug pipeline reports, debt-focused reports, etc.). I also remove reports that are associated with multiple firms. This reduces the number of reports to 2.4 million.

The 2.4 million reports are unevenly distributed across the sample period as shown in Figure 2.1.

The firm coverage grows steadily from approximately 250 firms in 1984 until it plateaus at around 1500 firms in 1999. Both the median number of reports per firm and the median number of contributors per firm increase almost monotonically over the sample period, with a slight drop in the median number of reports per firm in the mid-2000s. Both the number of reports and the number of contributors per firm are right-skewed. To avoid overweighting individual firms due to the right-skewed distribution of reports per firm, most of the paper will focus on firm-level aggregates instead of individual analyst opinions.

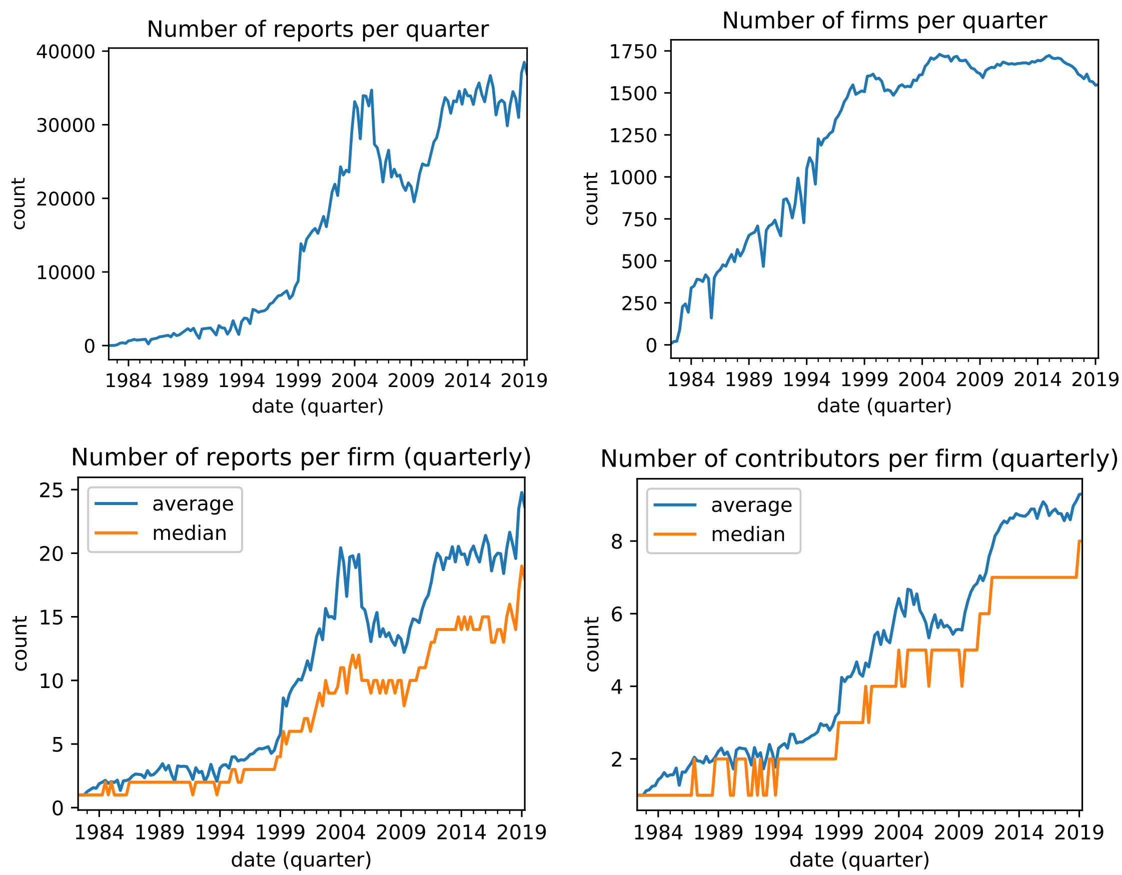
I match analyst reports to the numerical forecasts from I/B/E/S on the *estimator* level, corresponding to *contributor* in Thomson One. I obtain the full contributor names from the meta data of each report. I use a two-step process to match the contributor names to the numeric broker identifiers *emaskcd* and *estimator* in I/B/E/S. First, I use a text-search algorithm to extract target prices from the report texts and then match reports to the I/B/E/S target price database based on the date, firm, and target price. I use target prices since they are relatively easy to extract from the report text due to the fact that they are typically preceded by the term "target price" or "price target". Second, for contributors that cannot be linked through target prices - either because a contributor does not publish target price estimates, because our algorithm could not find the target prices on their reports, or because the contributor is not included in the I/B/E/S target price file - I hand-match reports by manually reading EPS forecasts from the reports and looking for matching entries in the I/B/E/S details file. The matching procedure is described in more detail in Section A.1.[7]
I limit the hand-matching to the largest 106 contributors by number of reports, accounting for 92.6% of the reports. 14 of these contributors could not be matched to I/B/E/S, neither via target price matching nor hand-matching. Together with 8 small contributors matched via target prices, I obtain a total of 100 valid contributor links that allow us to link 1.6m reports (75.0%) with estimates from the I/B/E/S details file.[8]

Text parsing

I strip the analyst reports of all parts that are unlikely to be economically meaningful, such as disclaimers and imprints, as well as all non-text information. In particular, I remove disclaimers, analyst certifications, legal notices, company profiles, report keys, figures, tables, and all other paragraphs that do not contain structured text. To reduce the size of the vocabulary, all words are lemmatized. More details on the pre-processing of the report text can be found in Appendix A.2.

[7] The matching procedure can also be used to match on analyst names. I choose to match on contributors since names might not be unique, thus matching on analyst names might be less accurate than matching on contributors.

[8] A report can be without an I/B/E/S link either because its contributor does not have a valid *emaskcd* link, or because there does not exist a matching firm-date-emaskcd entry in I/B/E/S.



The four panels show the quarterly composition of the cleaned analyst report sample. The sample spans Apr 1982 to Jun 2019 and consists of 2.4 million reports.

Figure 2.1: Analyst report sample composition

More details can be found in Section A.2.

Measuring tone in analyst reports

In natural language processing, there exists a trade-off between creating a sentiment measure that captures as much information as possible and creating a sentiment measure that can be easily interpreted. To maximize the interpretability of my findings, I use an established yet simple measure of textual sentiment to quantify the tone of the analyst reports, that is, a dictionary-based word-count model. Using the Loughran and McDonald (2011) sentiment dictionary for financial applications, I assign a naive tone score to each report. A report's tone is the difference between the number of positive words and negative words scaled by the total number of words in the report

$$Tone_{raw} = \frac{(\text{number of positive words}) - (\text{number of negative words})}{(\text{total number of words})} \tag{2.4}$$

A similar tone measure has been used by Gultekin et al. (2019) for analyst tone, as well as Hillert et al. (2014) to measure media tone and Schmeling and Wagner (2019) to measure central bank tone, among others. $Tone_{raw}$ is winsorized at the 1% level on either side. Summary statistics for $Tone_{raw}$ as well as its components are shown in Table 2.1 Panel B. Reports are on average 922 words long, with the report length being strongly right-skewed. On average, reports contain 24.68 positive words and 20.73 negative words, implying an average $Tone_{raw}$ of 0.37. The vast majority of words in each report are neither labeled positive nor negative, indicating that only a small part of the text corpus contains polarized information. For the remainder of the paper, I will use the standardized tone measure *Tone*, which is obtained by scaling $Tone_{raw}$ to mean zero and unit standard deviation.

Firm-level tone measures are constructed in close resemblance to consensus forecasts. Analysts frequently publish new reports without revising their estimates. For each analyst, firm, and earnings announcement, I select all reports within the 45-day window prior to the announcement date that follows the last revision, including the report on the revision date. The analyst-firm-date tone is calculated as the average *Tone* across the reports in the 45-day window. Finally, I average over analysts to obtain the firm-date tone $Tone_{i,t^-}$. Similarly, the post-announcement tone $Tone_{i,t}$ is obtained by first averaging the within-analyst tone in all reports linked to the latest forecast within 45 days of the announcement, and then averaging the analyst-firm-date tone across analysts.

2.2.4 Stock returns

Stock return data is from CRSP. Throughout the paper, stock returns are delisting-adjusted returns for common stock (share codes 10 and 11) traded on NYSE, Amex, and Nasdaq.

Panel A: Revisions and Errors

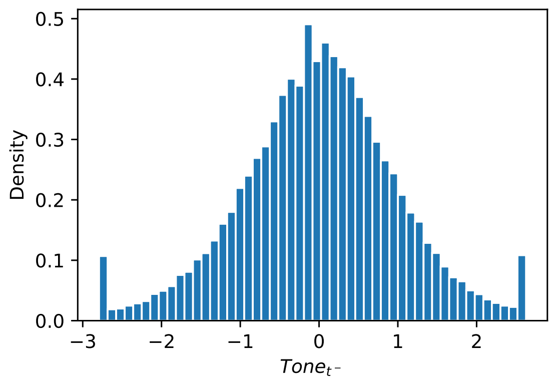
	count	mean	sd	min	p25	p50	p75	max
N_t^t	188436	0.00	1.00	−2.55	−0.45	−0.15	0.62	2.43
N_t^{t+1Q}	122647	0.00	1.00	−2.55	−0.61	0.09	0.62	2.44
N_t^{t+2Q}	104872	0.00	1.00	−2.55	−0.61	0.08	0.62	2.43
N_t^{t+3Q}	90911	0.00	1.00	−2.52	−0.62	0.09	0.62	2.42
N_t^{t+4Q}	71227	0.00	1.00	−2.38	−0.65	0.03	0.64	2.39
FE_{t-}^{t+1Q}	169236	0.00	1.00	−2.49	−0.55	−0.01	0.57	2.45
FE_{t-}^{t+2Q}	153218	0.00	1.00	−2.44	−0.58	0.10	0.56	2.46
FE_{t-}^{t+3Q}	137810	0.00	1.00	−2.44	−0.59	0.13	0.56	2.46
FE_{t-}^{t+4Q}	102826	0.00	1.00	−2.39	−0.61	0.12	0.59	2.39
FE_t^{t+1Q}	261437	0.00	1.00	−2.51	−0.50	−0.10	0.61	2.42
FE_t^{t+2Q}	227074	0.00	1.00	−2.41	−0.58	0.01	0.58	2.43
FE_t^{t+3Q}	207172	0.00	1.00	−2.37	−0.61	0.12	0.57	2.43
FE_t^{t+4Q}	186968	0.00	1.00	−2.36	−0.62	0.13	0.58	2.41

Panel B: Tone measure (report level)

	count	mean	sd	min	p25	p50	p75	max
n(pos)	1587084	24.68	29.87	0.00	9.00	17.00	30.00	2439.00
n(neg)	1587084	20.73	25.53	0.00	8.00	14.00	25.00	1637.00
n(total)	1587084	921.77	1036.03	1.00	411.00	682.00	1068.00	71265.00
$Tone_{raw}(\%)$	1587084	0.37	1.79	−4.47	−0.75	0.35	1.50	5.01
Tone	1587084	0.00	1.00	−2.71	−0.63	−0.01	0.63	2.60

Panel A shows descriptive statistics for earnings surprises, earnings revisions, pre- and post-announcement errors. All variables have been standardized to mean zero and unit standard deviation. Firm subscripts are omitted for ease of notation. Panel B shows descriptive statistics for the analyst tone measure *Tone* and its inputs for the sample of reports that have been matched with I/B/E/S. *n(pos)* is the number of positive words in a report, *n(neg)* the number of negative words, and *n(total)* the total number of words. $Tone_{raw}$ is the net number of positive words over the total number of words multiplied by 100. *Tone* is equal to $Tone_{raw}$ rescaled to mean zero and unit standard deviation.

Table 2.1: Summary statistics



The figure shows the distribution of shows the distribution of the firm-level pre-announcement analyst tone $Tone_{i,t^-}$ as defined in Section 2.2.3.

Figure 2.2: Distribution of analyst tone

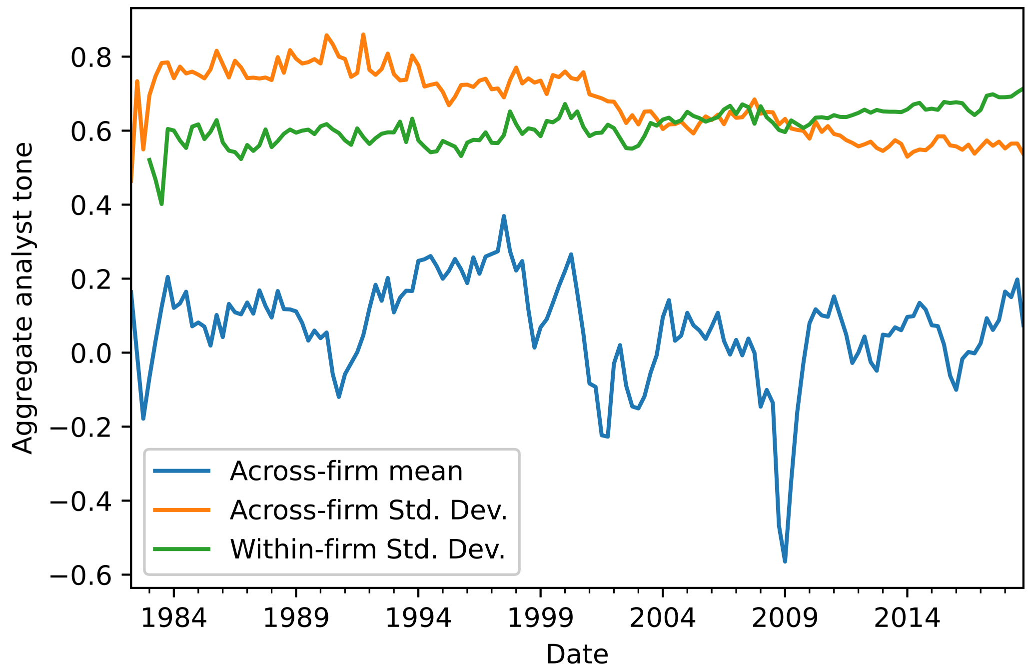
2.3 Properties of analyst tone

To gain first insights into the dynamics of analyst tone, I first investigate the evolution of quarterly analyst tone over the time. Figure Fig. 2.3 plots the across-firm mean, across and within-firm standard deviation from 1982 to 2019. Average analyst tone appears to be stationary with a high value during the dot-com bubble in the late 1990s and a low value during the 2008-09 financial crisis. The across firm standard deviation has been trending downward throughout the sample, from above 0.8 in the mid-1990s to below above 0.6 in 2019. Within-firm standard deviation has been relatively constant.

Next, I study the relationship between analyst tone and other measures of analyst opinion as well as firm characteristics. To allow for non-linearities, I focus on a non-parametric investigation of binscatter plots of analyst tone against various other variables. Even though the binscatter plots reveal that a linear model is clearly misspecified for many of the control variables, linear regression results can be found in Table A.7.

The top-left plot in Fig. 2.4 investigates the time-series properties of $Tone_{i,t^-}$. Analyst tone appears to be relatively short-lived, with the Pearson correlation between pre-announcement analyst tone from one quarter to the next one being 34%. The slope coefficient is around 2, suggesting that firm-level analyst tone decays by approximately half from quarter to quarter.

The top-right plot plots analyst tone against revisions of near-term earnings forecast between the current and the previous period. In particular, the horizontal axis plots $\frac{\mathbb{E}_{t-1}[C^t]-\mathbb{E}_{t^-}[C^t]}{P_{t-1}}$, the revision



This plot shows the evolution of the average analyst tone over time. Analyst tone is the textual measure *Tone* as defined in Eq. (2.4), standardized to mean zero and unit standard deviation. For each quarter, I first compute the average tone for each analyst-firm pair. For the across-firm mean, I average over analysts and then average over firms. For the across-firm standard deviation, I average over analysts and then take the standard deviation across firms. For the within-firm standard deviation, I take the standard deviation over analysts and then average across firms. Within-firm standard deviation is missing for the first two quarters in the sample since it requires at least two observations per firm, thus having higher coverage requirements than the other two measures.

Figure 2.3: Evolution of average analyst tone over time

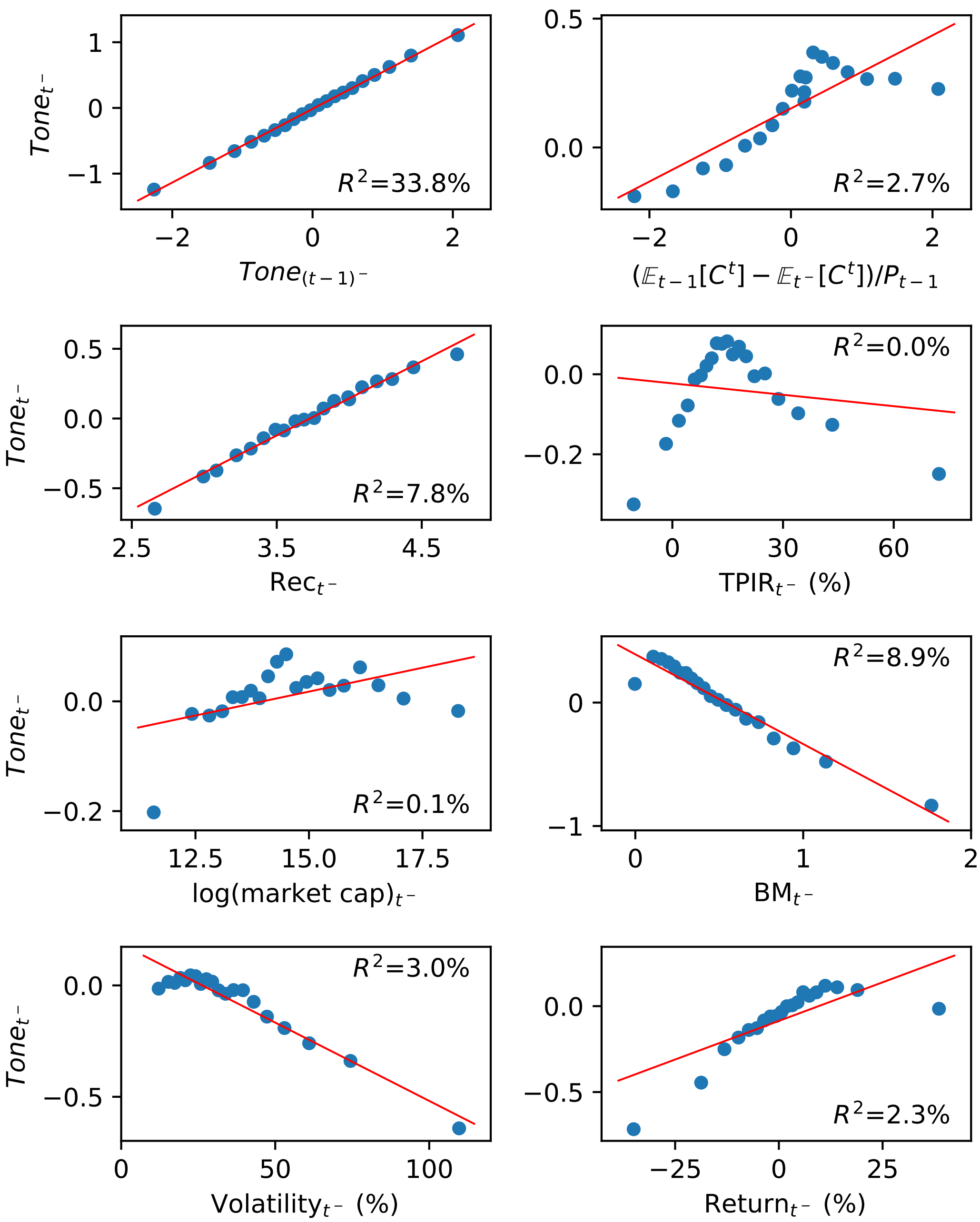
of period-t earnings expectations from shortly after the period-t-1 announcement to shortly before the period-t announcement. The plot reveals that there is no clear linear relationship between analyst tone and revisions of near-term earnings forecasts since the previous quarter. While negative revisions tend to be followed by negative analyst tone and positive revisions tend to be followed by positive analyst tone, tone is highest for marginally positive revisions. This suggests that analyst use the report text to describe or justify recent changes in their earnings estimates. An alternative reading of the plot is that analyst tone tends to be lower following larger revisions, an that this effect is stronger for negative than for positive revisions. The observed non-linear relationship between revisions and tone as well as the low R^2 of 2.7% suggest that analyst tone is likely to capture information beyond forecast revisions.

In addition to earnings forecasts, analysts frequently express their view on a firm via buy/sell recommendations and target prices. The two panels in the second row of Fig. 2.4 investigate the relationship between tone and these two opinion measures. The first panel reveals a linear relationship between analyst tone and buy/sell recommendations, which confirms that analyst tone is a meaningful measure of analyst opinions. An increase of Rec_{i,t^-} by one, which translates into an upgrade of the buy/sell rating by one category, predicts an increase of $Tone_{i,t^-}$ by approximately 0.5 standard deviations. However, the R^2 of this regression is just below 8%, suggesting that only a small fraction of analyst report tone can be explained by the buy/sell recommendation. The second panel shows an inverted-U shaped relationship between target price implied returns and analyst tone, with a peak at around 20%. There is no consensus in the literature as to whether high target price implied returns represent high expected risk premia or mispricing (Engelberg et al., 2019). While the interpretation of TPIR is beyond the scope of the paper, it is useful to note that there does not exist a clear linear relationship between TPIR and analyst tone.

Next, I study the relationship between analyst tone and firm characteristics. The third row of Fig. 2.4 plots analyst tone against size and the book-to-market equity ratio. Analyst tone appears to be fairly independent of size. However, I observe a strong linear relationship between book-to-market ratios and analyst tone: on average, analyst tone is substantially higher for growth firms. Firms with a book-to-market ratio below 0.5 tend to have positive analyst tone while firms with a book-to-market ratio above 0.5 tend to have negative analyst tone. The R^2 of a linear regression of analyst tone on book-to-market is 8.9%, suggesting that only a small fraction of analyst tone can be explained by the book-to-market ratio.

The fourth and last row of Fig. 2.4 plots analyst tone against the volatility and total return within the pre-announcement measurement window [t-45, t-1]. Average analyst tone is approximately constant across low to medium volatility firms, it is notably lower for firms in the 80th volatility percentile. This finding is in line with the observation that analyst tone is lower following large revisions in earnings forecasts, as new, publicly available information should trigger both forecast revisions and price changes.

Lastly, I observe a positive relationship between returns and analyst tone. Firms that had a positive return during the pre-announcement measurement window [t-45, t-1] tend to have positive analyst tone and returns that had a negative return negative analyst tone. I further investigate the causal relationship between analyst tone and stock prices in Section 2.4.1.



For each panel, the variable on the horizontal axis is sorted into 20 equally-sized bins. Each dot represents the mean analyst tone $Tone_{t^-}$ and the mean of the binned variable for one bin. The red line and the R^2 represent a linear fit of the underlying data. Firm subscripts are omitted for ease of notation. The first row plots analyst tone against one-quarter lagged analyst tone and revisions of current-quarter earnings expectation. The second row plots analyst tone against analyst recommendations and 12-months TPIR. The third row plots analyst tone against market capitalization and B/M. The last row plots analyst tone against the volatility and sum of daily returns during the tone measurement window.

Figure 2.4: Pre-announcement analyst tone and control variables

2.4 Earnings information in analyst reports

In this section, I investigate whether analyst tone contains information about future earnings. Huang et al. (2014) suggest that stock markets react positively upon the publication of analyst reports with positive textual opinion and vice versa controlling for revisions in near-term earnings forecasts. Here, I take an alternative approach and focus on analyst report publications that do not coincide with an update in earnings forecasts of any horizon, which allows us to isolate the reaction to analyst tone without imposing a functional form on the relationship between forecast revisions of various horizons and the associated market reaction. Figure 2.5 shows stock returns around the publication date of analyst reports that do come with a revision in earnings forecasts. Several observations stand out: First, I observe a positive market reaction to positive reports and a negative market reaction to the most negative reports. Interestingly, announcement returns are negative only for the lowest tone quintile. A possible explanation for this phenomenon is that analysts might be hesitant to publish negative opinions, for example in order to preserve banking relationships with their clients, which leads to an over-representation of positive reports in the sample. Such a publication bias would also explain the positive bias in analyst report tone documented in Section 2.2.3. Second, I observe a significant pre-publication drift for both positive and negative tone publications. Approximately half of the trailing 10-day cumulative abnormal return on the publication day is realized before the publication day. This observation suggests that analyst tone partially reflects information that has already been disseminated to the public and has been incorporated into stock prices before the publication of the report. The significant return on the publication date indicates that the report itself is informative: on average, reports in the highest tone quintile coincide with a 48bps increase in the stock price, while reports in the lowest tone quintile coincide with an 18bps decrease in the stock price. Third, I do not observe a significant post-publication drift. Stock prices seem to drift slightly on the first day after publication, however, this might be explained by the fact that some reports are published after regular trading hours on day 0. The lack of post-publication drift suggests that the earnings information transmitted through analyst tone is quickly incorporated into prices.

The announcement reactions suggest that analyst tone is either negatively correlated with discount rates or positively correlated with earnings expectations. If analyst tone contains no information about earnings expectations beyond the expectations disclosed in the numerical forecasts, earnings surprises should not be predictable by the analyst tone. To test this hypothesis, I regress realized forecast errors in period t on pre-announcement tone,

$$FE^{t}_{i,t^-} = \beta_1 Tone_{i,t^-} + \gamma Controls_{i,t} + \varepsilon_{i,t^-} \tag{2.5}$$

where $Controls_{i,t}$ is a matrix of fixed effects. Results are shown in Table 2.2 Column (1). I can confidently reject the hypothesis that analyst tone does not contain earnings information beyond the numerical earnings forecasts. A one standard deviation increase in analyst tone predicts a 0.04 standard deviations increase in period-t error. The effect is economically small, but of a similar magnitude as other predictors of forecast errors (see e.g. Bouchaud et al., 2019).

Next, I investigate whether analyst tone contains information about further-ahead earnings. Earnings tend to be autoregressive, so we expect some degree of predictability to follow mechanically from the predictability of current-period errors. To abstract from this mechanic correlation, I regress revisions of 4Q-ahead forecast errors on analyst tone while controlling for realized current-period earnings surprises,

$$FE_{i,t^-}^{t+4Q} = \beta_1 Tone_{i,t^-} + \beta_2 FE_{i,t^-}^{t} + \gamma Controls_{i,t} + \varepsilon_{i,t^-}. \tag{2.6}$$

Controlling for realized current-period earnings allows us to disentangle longer-term forecast errors from the permanent component of near-term forecast errors. The results are shown in Table 2.2 Columns (4) to (7). Pre-announcement analyst tone strongly predicts revisions of 4Q-ahead forecasts. Controlling for near-term errors in the long-term predictive regression does not significantly change the coefficient on analyst tone, suggesting that analyst tone is most informative about future cash flows that do not result from permanent near-term shocks. Again, while the magnitude appears small - a one standard deviation increase in $Tone_{t^-}$ predicts a 0.09 standard deviation error in the forecast for the next quarter - it is reasonably large compared to other predictor variables. The coefficient on $Tone_{t^-}$ ranges between 49% and 82% of the coefficient on $FE_{t^-}^{t}$, suggesting that a one standard deviation increase in analyst tone $Tone_{t^-}$ has about half the economic significance as a one standard deviation increase in the current period error $FE_{t^-}^{t}$, despite the fact that current-period error contains forward-looking information from the perspective of an analyst prior to the earnings announcement.

The evidence presented in this section suggests that analysts use the report text to convey information that has not yet been incorporated into their numerical forecasts, in particular about further-ahead fiscal periods. This is consistent with a model in which processing soft information is costly. In such a model, analysts might rationally choose not to update their estimates upon arrival of new information. If the processing of less precise signals is more costly, less precise information is less likely to be incorporated in the estimates. This is in line with the finding that analyst tone is particularly powerful in predicting revisions of further-ahead forecasts, as the signal-to-noise ratio of information about further-ahead earnings is likely to be lower than for current-period earnings.

I further explore this channel by studying revisions of long-term earnings forecasts around announcement dates. If analysts choose to transmit information about further-ahead earnings through

	$FE_{t^-}^{t}$	$FE_{t^-}^{t}$	$FE_{t^-}^{t}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$
$Tone_{t^-}$	0.057***		0.058***	0.104***		0.109***	0.102***
	(10.20)		(10.22)	(11.96)		(13.07)	(12.42)
Rec_{t^-}		0.025*	-0.006		0.002	-0.053**	-0.048**
		(1.90)	(-0.46)		(0.12)	(-2.60)	(-2.48)
$FE_{t^-}^{t}$							0.150***
							(21.94)
Fixed Effects	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)
N	66,531	66,531	66,531	34,359	34,359	34,359	34,359
R^2	0.131	0.129	0.131	0.188	0.182	0.188	0.208
Within R^2	0.002	0.000	0.002	0.008	0.000	0.008	0.032

This table reports regressions of pre-announcement forecast errors onto pre-announcement analyst tone. All variables are measured at the firm level, therefore, firm subscripts are omitted for ease of notation. Pre-announcement forecast errors $FE_{t^-}^{t+\tau}$ are realized actuals for period $t+\tau$ (shown in the superscript) minus consensus expectations measured prior to the earnings announcement in period t scaled by the stock price 46 days prior to the announcement. Pre-announcement consensus expectations are the average of individual analyst expectations measured within 45 days prior to the announcement. Pre-announcement analyst tone $Tone_{t^-}$ is the average tone of the reports published along with the selected individual expectations. See Section 2.2.3 for details on the measurement and aggregation of analyst tone. Regressions include firm (f) and year-quarter (t) fixed effects. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively.

Table 2.2: Pre-announcement tone and forecast errors

analyst tone instead of updating their numerical estimates due to the high uncertainty around the earnings in further-ahead periods, we expect the information in analyst tone to be gradually incorporated into the numerical estimates while we move closer to the earnings announcement day. To test this hypothesis, I test regress revision of long-term earnings forecasts onto analyst tone,

$$N_{i,t}^{t+\tau} = \beta_1 Tone_{i,t^-} + \beta_2 FE_{i,t}^{t} + \gamma Controls_{i,t} + \varepsilon_{i,t} \tag{2.7}$$

where τ again ranges from one to four quarters and $Controls_{i,t}$ is a matrix of fixed effects. Similar to Equation 2.4, I control for revisions in current-period forecasts (i.e. realized errors) to distinguish between the predictability of the permanent component in current-period earnings and the predictability of long-term earnings. Results are shown in Table 2.3. Pre-announcement analyst tone strongly predicts revisions in long-term earnings forecasts with a positive sign. As in the forecast error regressions, controlling for current-period revisions has virtually no impact on the coefficient estimate on longer-term revisions, lending further evidence to the hypothesis that analyst tone carries information about long-term earnings that is orthogonal to short-term earnings news. This information is at least partially incorporated into the numerical earnings forecasts upon arrival of new hard information on earnings announcement dates.

	N_t^{t+1Q}	N_t^{t+2Q}	N_t^{t+3Q}	N_t^{t+4Q}
$Tone_{t^-}$	0.112***	0.091***	0.079***	0.111***
	(15.86)	(12.27)	(11.14)	(12.53)
$FE_{t^-}^t$	0.191***	0.096***	0.085***	0.286***
	(23.53)	(11.26)	(7.48)	(27.08)
Fixed Effects	(f,t)	(f,t)	(f,t)	(f,t)
N	48,157	42,536	37,090	31,694
R^2	0.145	0.099	0.095	0.172
Within R^2	0.041	0.013	0.010	0.072

This table reports regressions of forecast revisions onto pre-announcement analyst tone. All variables are measured at the firm level, therefore, firm subscripts are omitted for ease of notation. Forecast revisions $N_t^{t+\tau}$ are changes in the consensus forecast for period $t+\tau$ (shown in the superscript) around the period t earnings announcement, measured as post-announcement consensus expectations minus pre-announcement consensus expectations scaled by the stock price 46 days prior to the announcement. Pre-announcement consensus expectations are the average of individual analyst expectations measured within 45 days prior to the announcement, and post-announcement consensus expectations are the average of individual analyst expectations measured within 45 days after the announcement. Pre-announcement analyst tone $Tone_{t^-}$ is the average tone of the reports published along with the selected individual pre-announcement expectations. See Section 2.2.3 for details on the measurement and aggregation of analyst tone. Regressions include firm (f) and year-quarter (t) fixed effects. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively. All variables are measured at the firm level, firm subscripts are omitted for ease of notation.

Table 2.3: Pre-announcement tone and forecast revisions

2.4.1 Is analyst tone priced?

Having shown that analyst tone contains information about future earnings, I now turn my attention to whether this information is reflected in market prices. Since analyst opinions are available to most institutional investors, any information in these reports should be quickly incorporated into market prices. The lack of post-publication drift shown in Section 2.5 suggests that markets quickly incorporate textual information. However, this observation by itself does not tell us whether markets do so in an unbiased way. If the market reaction to analyst tone is unbiased, earnings announcement returns should not be predictable by pre-announcement analyst tone. To test this hypothesis, I estimate

$$BHAR_{i,t}[0,2] = \beta Tone_{i,t^-} + \gamma Controls_{i,t} + \varepsilon_{i,t} \tag{2.8}$$

where BHAR$_{i,t}$[0,2] is the abnormal buy-and-hold return starting on the day of the announcement of period t earnings until the second trading day after the announcement. Abnormal returns are measured relative to a Fama-French three-factor model, with factor loadings calculated based on 90 daily returns in the window [-135,-46] prior to the publication. I do not consider firms with less than 45 valid returns in this window. $Controls_{i,t}$ is a matrix of fixed effects. Results are shown in Table 2.4 Column (1). I do not find a statistically significant relationship between pre-announcement analyst tone and announcement day returns.

Our analysis of forecast errors showed that analyst forecasts are understated for high analyst tone firms and overstated for low analyst tone firms. Since tone does not predict market reactions, this implies that N_t is misspecified as a measure of earnings surprise from the perspective of the market. Put differently, announcement day returns should appear too small relative to the observed revision in earnings expectations for high $Tone_{t^-}$ firms. This is confirmed in Table 2.4 Column (3). Controlling for the observed forecast revision, the sign on $Tone_{t^-}$ is negative, and the coefficient on the forecast revision itself increases when controlling for $Tone_{t^-}$.

Together, the findings in this section suggest that analyst tone contains information about future earnings above and beyond the information embedded in the same analysts' numerical earnings forecasts. Analyst tone appears to be particularly informative about further-ahead quarterly earnings. Stock returns on report publication days confirm this finding. However, analyst tone does not predict returns on earnings announcement days, suggesting that the information transmitted by analyst tone is adequately incorporated into prices upon publication of the reports.

2.5 Reaction to news

In this section, I investigate how the soft-processing of information prior to an announcement impacts the processing of new information arriving on the announcement date. In a model with costly information acquisition (e.g. Grossman and Stiglitz, 1980) or information processing constraints (e.g. Sims, 2003), the existence of directional soft-processed information can be informative in itself. If the net soft information is negative, it reveals that the analyst chose to acquire and soft-process more negative than positive information. This implies that analyst tone can be used as a measure of relative attention to negative versus positive events. For example, if analyst tone is nega-

	$\text{BHAR}_t[0,2]$	$\text{BHAR}_t[0,2]$	$\text{BHAR}_t[0,2]$
$Tone_{t^-}$	0.053		-0.253***
	(1.61)		(-6.54)
N_t^{t+1Q}		1.748***	1.776***
		(34.41)	(34.70)
Fixed Effects	(t)	(f,t)	(f,t)
N	61,745	39,308	39,308
R^2	0.055	0.134	0.135
Within R^2	0.000	0.064	0.065

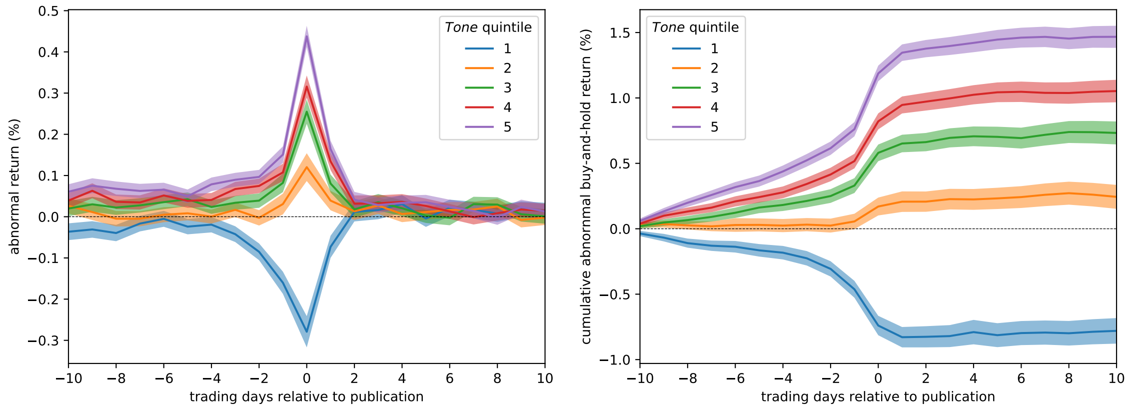
This table reports regressions of earnings announcement returns onto pre-announcement analyst tone. All variables are measured at the firm level, therefore, firm subscripts are omitted for ease of notation. BHAR_t[0,2] is the abnormal buy-and-hold return starting on the day of the announcement of period t earnings until the second trading day after the announcement. Abnormal returns are measured relative to a Fama-French three-factor model, with factor loadings calculated based on 90 daily returns in the window [-135,-46] prior to the publication. I do not consider firms with less than 45 valid returns in this window. N_t^{t+1Q} is the revision in 1Q-ahead earnings forecasts, measured as post-announcement consensus expectations minus pre-announcement consensus expectations scaled by the stock price 46 days prior to the announcement. Pre-announcement consensus expectations are the average of individual analyst expectations measured within 45 days prior to the announcement, and post-announcement consensus expectations are the average of individual analyst expectations measured within 45 days after the announcement. Pre-announcement analyst tone $Tone_{t^-}$ is the average tone of the reports published along with the selected individual pre-announcement expectations. See Section 2.2.3 for details on the measurement and aggregation of analyst tone. Regressions include firm (f) and year-quarter (t) fixed effects. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively. All variables are measured at the firm level, firm subscripts are omitted for ease of notation.

Table 2.4: Pre-announcement tone and announcement returns

tive, negative events are more salient than positive events. Moreover, the existence of soft-processed information reveals that the analyst chose to transmit this information in its soft form, rather than converting it into hard information. If converting soft to hard information is costly, this implies that this information is difficult or expensive to process relative to its usefulness for investors.

To establish a benchmark for the analysis in this section, suppose prices were following a martingale. The martingale hypothesis implies that the slope coefficient in a regression of long-run announcement returns onto short-run announcement returns is equal to one, i.e. $\beta = 1$ in

$$BHAR_{i,t}[0,90] = \beta BHAR_{i,t}[0,2] + \gamma Controls_{i,t} + \varepsilon_{i,t} \tag{2.9}$$



The figures show the daily (left panel) and cumulative (right panel) abnormal returns around the publication of analyst reports. Analyst reports are split into five equally sized groups based on their tone according to Eq. (2.4), where group 1 has the most negative tone and group 5 has the most positive tone. Day 0 is the date of publication. Abnormal returns are measured relative to a Fama-French three-factor model, with factor loadings calculated based on 90 daily returns in the window [-135,-46] prior to the publication. I do not consider firms with less than 45 valid returns in this window. To disentangle forecast revisions from analyst tone, I only consider reports that do not revise earnings forecasts and that are published at least 10 days and at most 90 days after the last earnings forecast revision of the same estimator. The sample consists of 342,485 reports. Shaded regions show 95% confidence intervals.

Figure 2.5: Stock market reaction to analyst reports

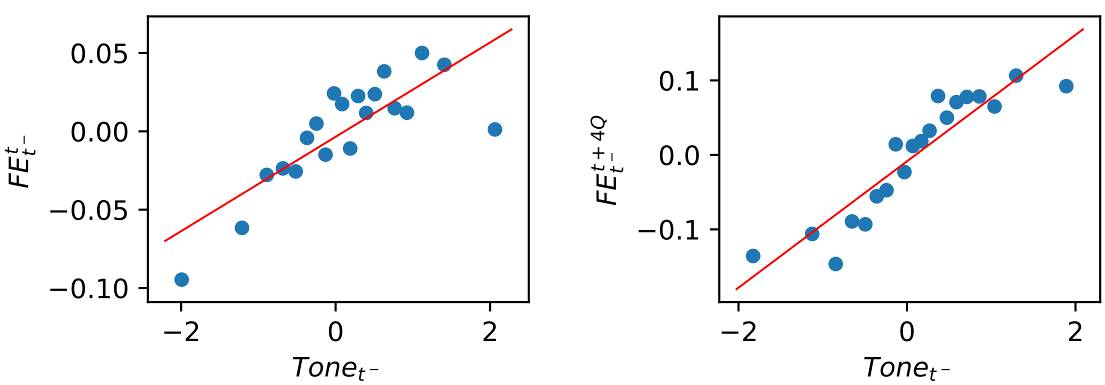
where $BHAR_{i,t}[0,T]$ is the abnormal buy and hold return of security i from t to T with respect to a Fama-French three-factor model. In contrast to the martingale hypothesis, a β above one indicates slow reaction to news, while a β below one indicates overreaction to news. In our sample of firms with valid pre-announcement tone $Tone_{t^-}$, the average β coefficient estimate after controlling for firm and time fixed effects is almost exactly equal to 1 as documented in Table 2.6 Column (1). In a similar setting, Martineau (2019) documents a β_1 of 1.1-1.2 for S&P 1500 firms in the post-2000 period, where most of our observations are located. I attribute the smaller slope coefficient estimate to the fact that firms with frequent analyst coverage tend to be larger firms with higher trading volume, where limits to arbitrage tend to be small and market efficiency tends to be high. Table 2.6 Column (2) estimates separate slope coefficients for firms with low, medium, or high pre-announcement analyst tone. None of the slope coefficients are statistically different from one.

In the previous section, I document that pre-announcement tone contains directional information about future earnings, and that this information is, on average, efficiently incorporated into

	BHAR[-10,10]	BHAR[-10,-1]	BHAR[0,10]
Tone	0.801***	0.435***	0.369***
	(16.00)	(14.27)	(13.68)
Constant	Yes	Yes	Yes
N	342,485	342,485	342,485
R^2	0.004	0.003	0.002

This table reports regressions of abnormal returns around analyst report publication on the analyst tone in a single report. BHAR[a,b] is the abnormal buy-and-hold return starting *a* days prior to the publication day and ending *b* after. Abnormal returns are measured relative to a Fama-French three-factor model, with factor loadings calculated based on 90 daily returns in the window [-135,-46] prior to the publication. I do not consider firms with less than 45 valid returns in this window. Analyst tone *Tone* is the analyst tone in the given report as defined in Section 2.2.3. Returns are shown in percent. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively.

Table 2.5: Stock market reaction to analyst report publication



The figure shows binscatter plots of current (left panel) and four quarter ahead (right panel) forecast errors against analyst tone. Forecast errors are relative to expectations measured in a 45-day window prior to the period-*t* announcement. Analyst tone is measured in the same time period. Forecast errors and analyst tone are defined in Section 2.2.1 and Eq. (2.4), respectively. For each panel, the analyst tone is sorted into 20 equally-sized bins. Each dot represents the mean analyst tone and the mean forecast error for one bin. The red line represents a linear fit of the underlying data. All variables are measured at the firm level, firm subscripts are omitted for ease of notation.

Figure 2.6: Analyst tone forecast errors

pre-announcement prices. Here, I investigate whether hard information about future earnings arriving on earnings announcement dates is processed differently conditional on whether or not it is in line with the information in analyst tone.

To start off, I conduct a graphical analysis of the post-earnings announcement drift *conditional on the sign of the news* and analyst tone tercile. The left panel of Figure 2.7 shows that following positive news, stocks with a low pre-announcement tone show significant drift in the 90 days following the announcement. The right panel of the same figure shows that following negative news, stocks with a low pre-announcement tone show significant overreaction that reverses within the first 15 days following the announcement.

To further quantify these results, I modify Equation 2.9 to estimate different slope coefficients conditional on the interaction between analyst tone and the sign of the news arriving on the earnings announcement date:

$$\begin{aligned} BHAR_{i,t}[0,90] = {} & \beta NewsSign_{i,t} \times ToneTercile_{i,t} \times BHAR_{i,t}[0,2] \\ & + \gamma Controls_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{2.10}$$

where *NewsSign* is a vector of dummy variables for the sign of the news and $ToneTercile_{i,t}$ is a vector of dummy variables that indicates the tercile of the pre-announcement analyst tone. Here, I directly use the announcement return $BHAR_{i,t}[0,2]$ as the measure of news rather than the earnings surprise. This allows us to abstract from any measurement error in earnings news caused by the predictability of earnings news via analyst tone that I documented in the previous section. In addition, the announcement day return captures both short-term and long-term earnings news without restricting our sample to firms for which I observe a revision in long-term earnings forecasts. This is particularly important given the evidence of the previous section that analyst tone is a stronger predictor of long-term earnings news than short-term earnings news. Results are shown in Table 2.6 Column (3). I find that markets overreact to news that is in line with pre-announcement analyst tone, e.g. negative news following a negative pre-announcement tone or positive news following a positive pre-announcement tone. In contrast, markets underreact to news that is not in line with pre-announcement analyst tone, e.g. positive news following a negative pre-announcement tone or negative news following a positive pre-announcement tone. The effect is slightly stronger for negative news than for positive news. The effect is economically large: for example, the coefficient estimate of 0.82 for negative news and low analyst tone suggests that markets on average overreact by about 18%. The coefficient estimate of 1.14 for negative news and high analyst tone suggests that markets on average underreact by about 14%.

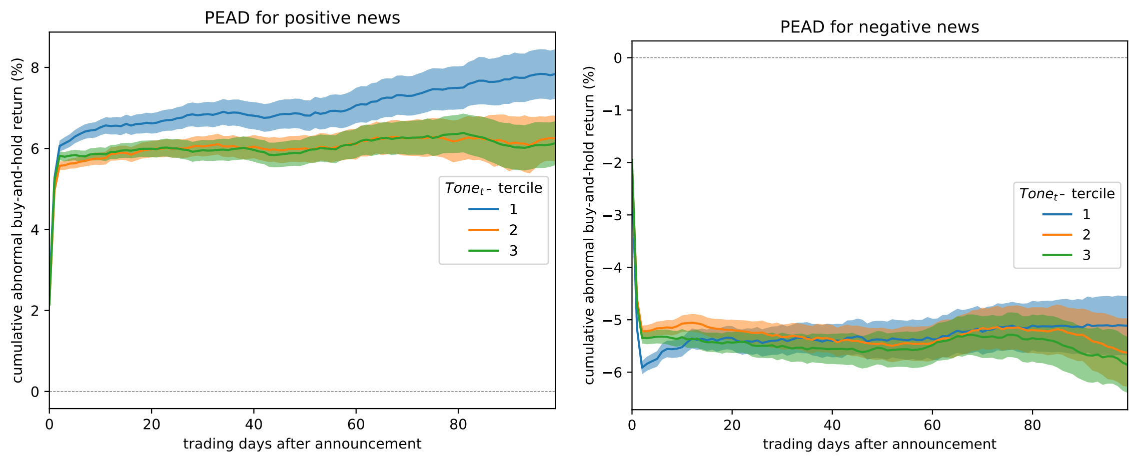
	BHAR[0,90]	BHAR[0,90]	BHAR[0,90]
BHAR[0,2]	1.0000		
	(0.0229)		
$1\{Tone_{t^-}$ low$\} \times$ BHAR[0,2]		0.9745	
		(0.0304)	
$1\{Tone_{t^-}$ med$\} \times$ BHAR[0,2]		1.0316	
		(0.0285)	
$1\{Tone_{t^-}$ high$\} \times$ BHAR[0,2]		0.9967	
		(0.0296)	
$1\{$BHAR[0,2] $< 0\} \times 1\{Tone_{t^-}$ low$\} \times$ BHAR[0,2]			0.8214***
			(0.0540)
$1\{$BHAR[0,2] $< 0\} \times 1\{Tone_{t^-}$ med$\} \times$ BHAR[0,2]			1.0798*
			(0.0410)
$1\{$BHAR[0,2] $< 0\} \times 1\{Tone_{t^-}$ high$\} \times$ BHAR[0,2]			1.1355***
			(0.0396)
$1\{$BHAR[0,2] $> 0\} \times 1\{Tone_{t^-}$ low$\} \times$ BHAR[0,2]			1.1085***
			(0.0393)
$1\{$BHAR[0,2] $> 0\} \times 1\{Tone_{t^-}$ med$\} \times$ BHAR[0,2]			0.9920
			(0.0425)
$1\{$BHAR[0,2] $> 0\} \times 1\{Tone_{t^-}$ high$\} \times$ BHAR[0,2]			0.8867***
			(0.0427)
Fixed Effects	(f,t)	(f,t)	(f,t)
N	61,028	61,028	61,028
R^2	0.184	0.184	0.185
Within R^2	0.113	0.113	0.115

The table reports coefficient estimates for variants of the following regression model:

$$BHAR_{i,t}[0,90] = \beta NewsSign_{i,t} \times ToneTercile_{i,t} \times BHAR_{i,t}[0,2] + \gamma Controls_{i,t} + \varepsilon_{i,t}$$

where BHAR[0,T] is the abnormal buy-and-hold return starting on day of the announcement of period t earnings until the Tth trading day after the announcement. *NewsSign* is a vector of two dummy variables indicating the sign of BHAR[0,2]: $1\{$BHAR[0,2] $< 0\}$ is a dummy variable that is equal to one if BHAR[0,2] is smaller than 0, and zero otherwise. $1\{$BHAR[0,2] $> 0\}$ is defined respectively. *ToneTercile* is a vector of dummy variables indicating the tercile of the pre-announcement sentiment: $1\{Tone_{t^-}$ low $\}$ is a dummy variable that is equal to one if pre-announcement sentiment $Tone_{t^-}$ is in the lowest tercile across the entire sample, and zero otherwise. Similarly, $Tone_{t^-}$ med and $Tone_{t^-}$ high denote the middle and highest tercile. Pre-announcement analyst tone $Tone_{t^-}$ is the average tone of the reports published along with the selected individual pre-announcement expectations. *Controls* are firm (f) and year-quarter (t) fixed effects. Standard errors are clustered by firm and year-quarter and are shown in parentheses. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively, testing the null hypothesis that the coefficients are equal to one. All variables are measured at the firm level, firm subscripts are omitted for ease of notation.

Table 2.6: PEAD conditional on news sign and sentiment



The left panel shows the post-earnings announcement drift of stock with a positive abnormal announcement day return, i.e. $BHAR_{i,t}[0,2] > 0$. The right panel shows the post-earnings announcement drift of stock with a negative abnormal announcement day return, i.e. $BHAR_{i,t}[0,2] < 0$. Day 0 is the date of the earnings announcement. Abnormal returns are measured relative to a Fama-French three-factor model, with factor loadings calculated based on 90 daily returns in the window [-135,-46] prior to the publication. I do not consider firms with less than 45 valid returns in this window. Stocks are sorted into three groups based on their pre-announcement analyst tone $Tone_{t^-}$, with group 1 being the group with the most negative tone and group 3 being the group with the most positive tone. The pre-announcement tone is the average tone across analysts prior to the earnings announcement for each firm. Details on the firm-level aggregation of the tone measure can be found in Section 2.2.3. Shaded regions show 95% confidence intervals.

Figure 2.7: Post earnings announcement drift and reversal

2.5.1 Trading strategy

The over- and underreaction results suggest that a simple trading strategy that buys firms with negative pre-announcement analyst tone is profitable shortly after earnings announcements. For both positive and negative news, firms with low analyst tone show abnormally high returns. Figure 2.7 suggests that these abnormal returns are concentrated in the first two weeks after the announcement, in particular for negative news. Therefore, a trading strategy must be implemented at a sub-monthly level. To reduce the number of portfolio rebalancing dates, I implement a comparatively conservative weekly trading strategy. At the end of every week t, I form three equal-weighted portfolios of firms that announced earnings between Thursday of week $t-1$ and Wednesday of week t based on their pre-announcement analyst tone. The two-day gap between the latest earnings announcement and the end of the week ensures that the portfolios do not capture any of the initial earnings announcement reaction. I then hold the portfolios for a full week until the end of Friday of week $t+1$. Earnings announcements are clustered in certain weeks of the year, leading to empty or poorly diversified portfolios in some weeks. To reduce the impact of idiosyncratic noise in the portfolios, I require at least five firms in each portfolio.[9] Table 2.7 shows the results. A trading strategy that buys low pre-announcement analyst tone stocks and sells high pre-announcement analyst tone stocks in the week after the announcement delivers an abnormal return of 38bps per week. Virtually all of the profits come from the long side of the portfolio. From June 1996 to Feb 2019, 522 weeks (42%) had valid low-high portfolio returns, i.e. at least five stocks in both the long and the short portfolio. While the weekly abnormal return of 38bps annualizes to 21.68%, an investor who invests in this trading strategy only earns 8.59% per year due to the large number of non-trading weeks.

Consistent with my observations regarding pre-publication returns in Figure 2.5, I find that firms with rather pessimistic tone have a low loading on the momentum factor and vice versa. It is worth noting that while the momentum factor is important to explain the variation in the low-high portfolio, excluding it from the regression decreases the estimate for the intercept by less than 5bps (see Table A.2 in the appendix).

The rightmost column of Table 2.7 documents returns for portfolios sorted on analyst tone throughout the entire sample, not restricting the universe of firms to companies with earnings announcements in week $t-1$. I find that there is no significant abnormal return related to analyst tone using the full sample. This finding is consistent with the hypothesis that analyst tone only enters realized returns through the processing of additional, material earnings information, as in a model of limited investor attention.

Table 2.8 shows the return of low minus high analyst tone portfolios for different firm sizes. Firms in the lower two size terciles earn a weekly alpha of 33-35bps, close to the full-sample alpha of 38bps. The alpha decreases significantly for firms in the highest size tercile to 20bps per week. This finding supports the hypothesis that the abnormal returns documented in this section are due to limited attention or limits to arbitrage, which are likely to be more severe for small firms.

[9]The results remain virtually unchanged for a larger minimum requirement of firms, e.g. 10 or 20. Setting the threshold to 5 maximizes the number of weekly observations while keeping the idiosyncratic portfolio risk moderate.

	post-announcement weeks				all weeks
	low	med	high	low-high	low-high
Intercept	0.354***	0.159***	-0.046	0.378***	-0.024
	(5.16)	(2.63)	(-0.86)	(4.58)	(-1.19)
Mkt-RF	1.014***	1.010***	1.056***	-0.053	0.004
	(34.68)	(36.20)	(42.71)	(-1.43)	(0.45)
SMB	0.502***	0.296***	0.386***	0.168***	0.007
	(9.58)	(6.60)	(9.19)	(2.64)	(0.49)
HML	0.277***	0.232***	0.143***	0.111*	0.203***
	(5.25)	(5.00)	(3.37)	(1.73)	(13.39)
Mom	-0.285***	-0.037	0.094***	-0.388***	-0.316***
	(-8.45)	(-1.20)	(3.36)	(-9.40)	(-32.28)
N	591	585	662	544	973
$Adj.R^2$	0.78	0.75	0.77	0.18	0.65

The table shows weekly risk-adjusted returns from an analyst tone-based trading strategy. The regressions use the Fama-French (1993) and Carhart (1997) four-factor model. The first four columns restrict the universe of stocks to firms that had an earnings announcement in the previous week. Every week, I form portfolios of stocks that announced earnings in the previous week based on their pre-announcement analyst tone. The *low*, *med*, and *high* portfolios contain firms with a pre-announcement analyst tone in the lowest, middle, and highest tercile relative to the entire sample of pre-announcement analyst tone, respectively. See Section 2.2.3 for details on the measurement of firm-level analyst tone. Firms have equal weights within each portfolio. I exclude weekly observations in which less than five firms were allocated to each portfolio. *low-high* is a zero-cost portfolio that buys the *low* portfolio and sells the *high portfolio* in weeks where both the *low* and the *high* portfolio have a valid return observation. *N* shows the number of trading weeks. The rightmost column shows the returns of a similar strategy that is not limited to firms that announced earnings in the previous week. The sample period is Feb 1985-Feb 2019, however, 95% of observations are after Jun 1996. Alphas are shown as weekly percent returns. T-statistics are shown in parentheses. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively.

Table 2.7: Risk-adjusted returns of an analyst tone-based trading strategy

	small	medium	big
Intercept	0.346**	0.333**	0.197**
	(2.13)	(2.48)	(2.38)
Mkt-RF	0.015	-0.018	-0.043
	(0.19)	(-0.29)	(-1.17)
SMB	-0.023	0.168	-0.058
	(-0.16)	(1.58)	(-0.93)
HML	0.162	0.213**	0.066
	(1.14)	(2.10)	(1.01)
Mom	-0.352***	-0.388***	-0.401***
	(-4.06)	(-5.79)	(-9.75)
N	237	289	370
$Adj.R^2$	0.14	0.17	0.25

This table shows returns of the low-high portfolio returns of a tone-based trading strategy similar to the post-announcement strategy in Table 2.7. Firms are split into terciles based on their market capitalization using NYSE breakpoints, labeled small, medium, and big.

Table 2.8: FF4 risk-adjusted returns of an analyst tone-based trading strategy, double-sorted on size

2.6 Conclusion

This chapter shows that analyst report texts contain valuable information about future earnings. A simple dictionary-based analyst tone measure predicts earnings several periods ahead above and beyond the numerical estimates issued by the same analysts. Analyst tone is particularly informative about longer-term earnings. Forecast revisions between the publication of an analyst report and the announcement of the actual earnings reveal that analysts partially incorporate the information previously transmitted through the report text into their numerical estimates over time. Stock prices react quickly to incorporate the earnings information in analyst tone into prices.

Pre-announcement analyst tone predicts post-earnings announcement drift patterns. If analyst tone is low, i.e. analyst reports focus on negative rather than positive information, markets overreact to negative news and underreact to positive news. I suggest that analyst attention to positive versus negative news can be used as a proxy for investor attention.

Chapter 3

Learning (from) the Market's Mistakes

3.1 Introduction

Do market prices efficiently incorporate all publicly available information? Since Fama (1965) formulated the efficient market hypothesis, a vast number of studies have attempted to test the semi-strong market efficiency hypothesis with mixed results. Empirical tests of this hypothesis face three main challenges. 1) How can we identify the information that is potentially not fully priced in? 2) How can we mitigate data mining concerns and ensure the out-of-sample validity of our findings? 3) How can we distinguish the unraveling of biased expectations from heterogeneous discount rates?

In this paper, I develop a hierarchical attention-based neural network (HAN) that allows me to identify and quantify textual information that is not reflected in asset prices. I train the HAN model to predict price changes on earnings announcement dates using sell-side analyst reports that were published in the preceding quarter. I demonstrate that earnings announcement returns are predictable ex-ante, i.e. before the release of the earnings numbers, using information that has been published weeks before the announcement. I show that capturing contextual information is crucial to extracting the pricing-relevant information from text. Furthermore, I propose a method to optimally aggregate information from multiple text documents to form a single asset return prediction.

The time sequence of report publications and announcements allows me to separate initial reactions to the publication of news, gradual information diffusion, and reaction to the announcement of the actual earnings. I document that markets initially react insufficiently to certain information in analyst reports. Prices show no sign of convergence after the initial reaction as one would expect from a model of gradual information diffusion. Instead, the insufficient initial reaction is persistent and is only corrected on the next earnings announcement date, giving rise to strong predictability of earnings announcement returns.

In the absence of risk premiums, returns on announcement days reflect cash flow news to the investor. If cash flow expectations are unbiased, cash flow news should be unpredictable given the information set of the investor. On the other hand, if cash flow expectations are biased, these biases are likely to be unraveled on the earnings announcement day when investors learn the true cash flow to equity holders. This unraveling of biases then generates predictability in announcement day returns. I conduct a series of tests to rule out that the announcement predictions capture announcement-day specific risk factors instead of biased cash flow expectations. I show that high announcement prediction stocks tend to have slightly lower market beta on both announcement and non-announcement days and I do not find that the announcement return predictions predict changes in factor betas. I am ruling out a learning-based channel as suggested by Savor and Wilson (2016) by showing that both firms with a positive and firms with a negative announcement return predictions have a positive correlation with future earnings growth. Instead, analyst forecasts are too low for announcements with a positive announcement return prediction and too high for announcements with a negative return prediction, which lends further support to a biased expectations explanation.

The large sample allows for effective cross-validation and out-of-sample testing of the identified prediction patterns. Randomly sampled historical time periods appear to be sufficiently independent to serve as a validation set for hyperparameter tuning. Not only does this allow me to increase the length of the out-of-sample period, but also to use very recent data for training and to study the stability of predictive relationships over time. I show that the predictors identified by the neural network persist several years out-of-sample but that the predictive power eventually subsides. This finding is in line with the idea that investors learn over time and that arbitrage opportunities gradually disappear. However, learning does not happen instantaneously, but over the course of several years.

The predictability of announcement day returns is connected to the cross-sectional predictability of returns at lower frequency. A portfolio sorted on out-of-sample announcement return predictions earns a CAPM alpha of 93bps per month, with the majority of the alpha being realized on earnings announcement days. The announcement prediction-sorted portfolio loads significantly on several well-known risk factors, such as value, momentum, and profitability. A six-factor model (Fama and French, 2015, plus momentum) leaves 49bps of the monthly alpha unexplained. It is worth emphasizing that the prediction model is not supplied with any form of numerical information about the firms. Nonetheless, the model picks up cross-sectional factors that are typically constructed from numerical firm characteristics such as past returns or accounting data. Alpha and most risk factor loadings increase monotonically with the predicted return for the next month. In particular, the factor model regressions reveal a strong relationship between the announcement predictions and the profitability factor. While the profitability factor can explain less than half of the alpha of the announcement prediction-sorted portfolio and the remaining alpha stays highly significant, the announcement prediction-sorted portfolio can explain two-thirds of the alpha on the profitability-sorted portfolio in the 2004-2019 period, with the remainder being statistically insignificant at the 10% level. This finding suggests that biases in cash flow expectations play an important role in explaining the returns on the profitability factor.

The neural network presented in this paper is inspired by similar networks that have been used in various machine learning applications, in particular Yang et al. (2016), who use a hierarchical attention network for document classification. The neural architecture is motivated by the sequential and hierarchical flow of textual data. Each earnings announcement return is matched with a set of documents (i.e. analyst reports), each of which contains a sequence of words. The HAN processes a set of documents in a series of annotation and aggregation steps. In annotation steps, data points are enriched by contextual information of the surrounding data points. Here, data points refer to either individual words or entire text documents. Annotation steps only impose a minimal amount of structure on the interactions between input variables and are largely motivated by our general understanding of linguistics and optimization intricacies of neural networks. In aggregation steps, the dimensionality of the input space is reduced by aggregating over a set of data points. The attention mechanism used in the aggregation steps allows us to aggregate parts of the text by using very flexible yet affine combinations. In addition to the potential performance gains[1], the affine structure allows us to derive economic insights from an otherwise opaque empirical model.

I benchmark the HAN against popular bag-of-words prediction approaches, most notably an Elastic Net. The HAN outperforms the Elastic Net by a factor of four. The key difference between the HAN and any bag-of-word approach is the HAN's ability to model interactions. In linguistics, interactions between words or text segments are often referred to as context. While a human reader would have immense difficulties comprehending a document article if it was presented as an unordered list of the words that comprise the document, much of the existing finance literature that utilizes textual analysis uses such a bag-of-words approach to extract the information content of the text. Bag-of-word methods remain the method of choice in economic research due to their relatively low computational needs and seemingly better interpretability. However, interpretability in bag-of-words models is largely by assumption. The interpretation that the authors typically attach to individual words in such an analysis is only valid as long as the context that we abstracted from in the analysis does not have a first-order impact on the word meaning, just as the interpretation of an OLS coefficient is only valid as long as the true underlying model is approximately linear. Therefore, I argue that the merits of the methodology presented in this paper easily outweigh its

[1] Attention mechanisms were initially developed as a tool to improve the performance of neural machine translation algorithms.

higher computational costs and slightly more difficult inference. In addition, I demonstrate how we can obtain interpretable explanations for the predictions of more complex prediction models such as the HAN under the assumption of local linearity.

My work contributes to the empirical literature on market efficiency and stock return predictability in the cross-section. Several papers have suggested that cross-sectional differences in expected returns can be attributed to biased cash flow expectations. Nagel (2005) shows that cross-sectional anomalies tend to be strongest among short-sale constrained stocks, suggesting that anomalies originate from mispricings that are hard to arbitrage. Lochstoer and Tetlock (2020) suggest that anomaly returns are driven by cash flow news. Engelberg et al. (2018) show that anomaly returns are six times higher on earnings announcement days compared to non-announcement days. Bouchaud et al. (2019) find that the profitability anomaly is concentrated among firms with sticky analyst expectations.

I also add to the broader literature on how market participants form earnings expectations. Ball and Brown (1968) and Bernard and Thomas (1989, 1990) show that markets underreact to earnings news. More recently, Chang et al. (2017) provide evidence that markets fail to price information in seasonal earnings patterns, leading to predictable market reactions to seasonal earnings.

In addition, I contribute to the emerging literature on machine learning in finance. A growing number of papers use machine learning techniques to capture non-linear relationships in the data and to tackle dimensionality reduction tasks in asset pricing. Bryzgalova et al. (2019) and Feng et al. (2020a) use machine learning tools to select factors and asset pricing models. Heaton et al. (2017), Feng et al. (2020b), and Gu et al. (2020) use feedforward networks to construct optimal cross-sectional return predictions. Chen et al. (2020) use a general adversarial network to learn the stochastic discount factor. All of the aforementioned papers differ from my work in two key aspects. First, they typically focus on the prediction of monthly stock returns while I focus on the prediction of returns on earnings announcement days. Second, their input space typically consists of a relatively small number of pre-defined covariates such as firm characteristics and past returns. Many of these covariates have been identified as predictors of the cross-section of returns by traditional research.[2] This means their neural networks face a substantially easier task than my network: while my network finds predictors from a relatively unstructured dataset from scratch, a network that is supplied with pre-defined predictor variables only needs to find optimal combinations of these variables. Moreover, when using pre-defined covariates that were only recently identified as predictors of returns, for example operating profitability, researchers introduce a look-ahead bias which inflates the out-of-sample R^2 values as they are no longer truly out-of-sample.

[2]Most papers use around 100 covariates or less. To put this number in perspective, consider that Hou et al. (2020a) document 452 anomalies and that Compustat alone has 974 variables.

Other applications of machine learning in asset pricing include Bali et al. (2020), who take the methodology of Gu et al. (2020) to the cross-section of corporate bond returns and Bianchi et al. (2021), who use machine learning tools to predict treasury bond returns from macroeconomic variables. van Binsbergen et al. (2020) use machine learning to de-bias analyst expectations.
Lastly, my paper extends the literature on natural language processing in finance. Previous research has documented the information content of various textual data sources. Antweiler and Frank (2004) and Tetlock (2007) provide evidence of the information content of internet stock message boards and news paper articles. Ke et al. (2020) show that newspaper sentiment predicts stock returns. Jiang et al. (2019) and Azimi and Agrawal (2021) show that the sentiment in corporate disclosures predicts stock returns. Huang et al. (2014) document the information content of analyst report texts. Gentzkow et al. (2019) provide an excellent overview of additional NLP applications in finance.

The rest of the paper is organized as follows. Section 3.2 introduces the methodology and reports in-sample results. Section 3.3 discusses out-of-sample results. Section 3.4 discusses slow information diffusion and investigates how investors learn from their past mistakes. Section 3.5 relates the results to the broader cross-section of stock returns. Section 3.6 discusses risk-based explanations for the findings. Section 3.7 offers some insights into the inner workings of the neural network and the underlying drivers of the return predictability. Section 3.8 concludes.

3.2 Methodology

3.2.1 Data

I obtain analyst reports from Thomson One. The sample contains 3.5 million reports and covers the period from 1982 to 2019. Due to low coverage and survivorship bias concerns for the first part of the sample, I only use the 2004-2019 sample for out-of-sample return predictions. I remove all tables, figures, footnotes, and disclaimers from the reports. Stock price data is from CRSP, and earnings announcement dates are from Compustat.

3.2.2 Preliminary considerations

Predicting returns from textual data means finding a mapping from a very large multidimensional space to a scalar. While any prediction task can be formalized as finding a mapping between inputs and outputs, it is particularly useful to think about a prediction task with textual inputs as finding a mapping between inputs and outputs due to the complexity of the input space. The algorithm that finds this mapping has to fulfill multiple criteria. First, it has to be able to process a high-dimensional and sparse input space (approximately 250,000 unique words). Second, it needs to capture interactions and dependencies between two or more words to stand a chance to capture a meaningful fraction of the information content in the report texts. Third, it has to be highly scalable to be able to process a large amount of data. Fourth, in order to provide economic insights beyond pure return predictability, it has to generate a mapping that is interpretable.

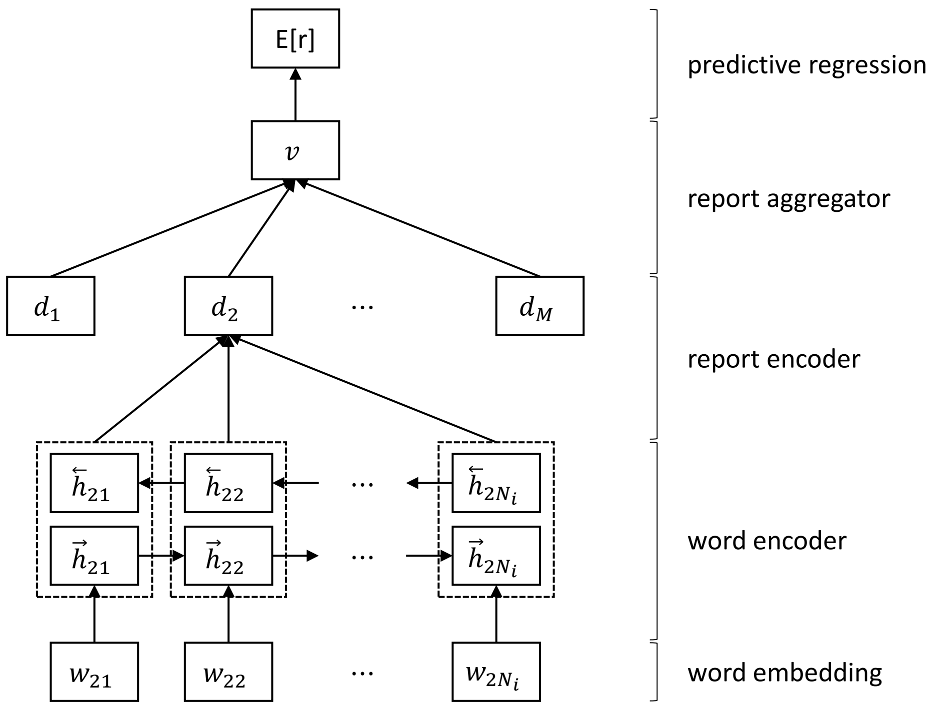
Textual data differs from the numerical panel data that is commonly used in finance and economics for two key reasons. First, text is sequential. While we can look up individual word meanings in a dictionary, the full information content of a word is only revealed when the word is considered in conjunction with the surrounding words. Second, text is hierarchical. For example, an academic journal is divided into papers, papers are divided into sections, sections are divided into sentences, and sentences are divided into words. In the context of analyst reports and earnings announcements, the textual data for a given firm-announcement can be divided into individual documents (i.e. the reports). The structure of each individual document varies, but the documents can always be divided into words.

Neural networks are a powerful tool to model complex functional mappings between complex input and output spaces. In fact, a standard feedforward network can approximate any continuous functional mapping between two Euclidian spaces (Hornik et al., 1989). In computational linguistics, the class of recurrent neural networks has emerged as the go-to tool to model the sequential structure of text. In contrast to a feedforward network, nodes in a recurrent neural network form a directed graph, which mirrors the way a human reader reads a text document. In this paper, I propose to use a deep recurrent neural network to estimate a functional mapping between analyst reports and announcement day returns.

To formalize the dimensionality reduction problem for a single announcement, let M be the number of reports, N be the number of words in each report, and V be the size of the vocabulary. The input space is of size $M \times N \times V$. Note that the last dimension is extremely sparse as it represents a one-hot encoded vector that identifies the word in the vocabulary. The inherent hierarchy of the data gives us a natural structure to approach this dimensionality reduction problem. First, I reduce the last dimension to a much smaller, dense representation $M \times N \times E$ where E is small, in particular, $E \ll V$. Next, I summarize or *encode* the information of an entire report in a single vector, resulting in a feature space of dimensions $M \times D$. D is chosen such that $D \ll N \cdot E$, so that the report encoding is significantly smaller than the previous representation. Next, I encode the information across reports in a single firm vector, resulting in a feature space of dimensions $F \times 1$. Finally, this vector is transformed into a scalar which represents the return prediction.

3.2.3 Neural architecture

The hierarchical structure of the neural network is visualized in Fig. 3.1 and discussed in detail in the following. To ease notation, I describe the algorithm in vector notation for a single announcement example and leave out firm-time subscripts. Let a given announcement be associated with M reports of length N_i, where $i \in [1, M]$ denotes a report. The algorithm in this paper ignores the temporal order of the reports. In other words, the order of the reports is treated as arbitrary. ω_{it} with $t \in [1, N_i]$ represents the tth word in the ith report. Word order is kept intact and will be taken into account by the algorithm. Each word ω_{it} is a one-hot encoded vector of length V, the size of the vocabulary.



The figure visualizes the architecture of the neural network for a single example. An example is a single firm-month, consisting of M reports of length N_i. Boxes represent vectors and arrows represent algebraic operations. In particular, w_{21} is the word vector (embedding) of the first word in the second report. h_{21} is the hidden state of the word encoder at the same position and is comprised of the forward hidden state $\overrightarrow{h}_{21}$ and the backward hidden state $\overleftarrow{h}_{21}$. d_2 is the report vector of the second report. v summarizes the information of all reports for the given firm-month. $\mathbb{E}[r]$ is a scalar representing the predicted return.

Figure 3.1: Hierarchical network architecture

Embedding

Given a report i with words $\omega_{i1}, \ldots, \omega_{iN_i}$, I first embed the words to low-dimensional vectors using an embedding matrix W_e. Word embeddings have increasingly been used in the finance literature to generate low dimensional word representations (see e.g. Gentzkow et al., 2019). These vector representations can be either learned jointly with the rest of the model or in a separate procedure that we call pre-training. For the task at hand, pre-trained *fasttext* embeddings turn out to be the most promising representation. See Section 3.2.4 for a further discussion. The embedded words are represented by vectors w_{it},

$$\underset{E\times 1}{w_{it}} = W_e \underset{V\times 1}{\omega_{it}}.$$

Word encoder

The next step is to annotate words with contextual information from the rest of the report. Simpler NLP approaches often model context by augmenting the feature representation by word combinations. This approach is either very restrictive with respect to the distance between words, i.e. by restricting the context to neighboring words as in the n-gram approach, or results in an exponentially growing feature space when considering interactions between all words in the document. In this paper, I take a different approach that utilizes the chain-like structure of textual data. The HAN reads the text sequentially, i.e. word-by-word, and maintains an internal state vector at every step. The state vector summarizes the contextual information of all previous words. At each step, the algorithm combines the current word vector with the internal state vector to produce a new, context-enriched word representation, which also serves as the updated state vector for the next word. This type of recurrent neural architecture has been proven to be highly successful in many natural language processing applications from classification (Schuster and Paliwal, 1997) to machine translation (Cho et al., 2014b).

In particular, I use a bidirectional gated recurrent unit (GRU, Cho et al., 2014a,b) to summarize information of both the preceding and succeeding words and to incorporate this information into the word annotations. Like any recurrent neural network, the GRU reads examples (reports) sequentially during the training process. At each step (that is, at each word) the network combines the information in the current word (represented by its word embedding) with the contextual information of all previous words in the same example. The contextual information is captured in a single hidden state vector that is updated at each step. A detailed description of the GRU including its algebraic representation can be found in Appendix B.1.2.

The bidirectional GRU walks through the report text twice, once in the natural order and once in reverse order. Using a bidirectional architecture has been shown to generate better representations of the linguistic context in a wide range of applications (Schuster and Paliwal, 1997). Letting $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ denote the vector-valued transformation functions of the forward and backward pass, respectively, the hidden states are given by

$$\underset{H\times 1}{\overrightarrow{h_{it}}} = \overrightarrow{GRU}(w_{it}, w_{it-1}, \ldots, w_{i1})$$
$$\underset{H\times 1}{\overleftarrow{h_{it}}} = \overleftarrow{GRU}(w_{it}, w_{it+1}, \ldots, w_{iN_i})$$

The total word annotation for a given word ω_{it} is obtained by concatenating the forward and backward hidden states

$$\underset{2H\times 1}{h_{it}} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}].$$

These concatenated hidden states can be interpreted as word embeddings that have been augmented by the contextual information of the entire document. Importantly, the hidden state takes into account positional information, allowing a recurring term to be represented by completely different vectors within a single report.

Report encoder

The next step is to aggregate the information of the word encodings $h_{i1}, \ldots, h_{it}$ of a given report into a single document vector d_i.

RNNs are often set up to simply pass the last hidden state of the recurrent layer to subsequent layers, effectively treating the last hidden state as a summary of the entire document. In this paper, I use a self-attention mechanism (Bahdanau et al., 2014; Luong et al., 2015) stacked on top of the recurrent layer to obtain optimal document representations. The attention mechanism constructs the document vector as a weighted sum of the hidden states of the recurrent layer. The attention weights are a function of the input data which is learned together with the rest of the model. Importantly, the weight of a single word is a function of all words in the document and its position within the document. This means any given term can receive vastly different weights even with a single document depending on the context of each instance of the term.

The implementation of the attention mechanism in this paper follows Yang et al. (2016). I first obtain a hidden representation u_{it} of the report content by feeding the hidden states h_{it} through a single-layer multi-layer perceptron,

$$\underset{2H\times 1}{u_{it}} = \mathsf{tanh}(W_w h_{it} + b_w).$$

Throughout the paper, I use the letter W to denote weight matrices and the letter b to denote bias vectors.[3] Weights are then calculated by comparing each transformed hidden state u_{it} with a trainable context vector u_w. Specifically, each hidden state's weight is calculated by measuring the alignment of the transformed hidden state with the context vector via the dot product and converting it to a non-negative weight with the softmax function,

$$\alpha_{it} = \frac{\mathsf{exp}(u_{it}^\top u_w)}{\sum_{t'} \mathsf{exp}(u_{it'}^\top u_w)}.$$

The context vector u_w can be thought of as the embedding of a fictitious word that is informative for return prediction and is learned in the training process. Word-level hidden states that are similar to this fictitious word tend to be more informative for return prediction, therefore they receive a higher weight in the report encoding.

The affine structure of the attention layer allows us to investigate which part of an input sequence drives the model prediction. Fig. 3.2 visualizes the attention weights α_{it} for a randomly selected analyst report. The neural network attends to various parts of the text, in particular forward-looking statements as well as discussions of prices and valuations.

After obtaining the attention weights we can calculate the aggregate hidden state as

$$\underset{2H\times 1}{\tilde{h}_i} = s_i^h \sum_t \alpha_{it} h_{it}.$$

where

$$s_i^h = 1 + \delta_h (N_i - 1)$$

is a scaling factor that is proportional to the number of words in the report N. The scaling factor allows the network to learn adjusted weights $s_i^h \alpha_{it}$ that do not add up to one. For example, $\delta_h = 1$ implies that the information within a report is additive, while $\delta_h = 0$ implies that the information in the report is best represented by the weighted average of the word-level vectors. Training the weights and the scaling factor separately, rather than directly learning weights that do not need to add up to one, allows me to directly observe the optimal scaling in the trained model. To obtain the

[3]In machine learning, a vector of constants is typically referred to as bias vector in analogy to biasing in electronics.

> we are resuming research coverage of altera corporation with a strong buy recommendation . as the semiconductor industry recovers from a three year downturn , altera is benefiting from a favorable pricing environment , strength in the communications sector and an improving cost profile that is allowing gross margins to expand . altera continues to impress us with the scope and breadth of its new product initiatives . it is apparent that the company is firing on all cylinders and laying the foundation necessary to maintain and extend a leading position in the programmable logic device (pld) marketplace . new products continue to firmly drive growth at the company while strong end markets and a greater mix of products and reduced pricing pressure are accelerating altera 's top line growth . we expect continued focus on cost reductions should provide ample opportunity for margin expansion . we anticipate that the pld industry should grow NUM in NUM and NUM . this exceeds consensus expectations of another NUM NUM in the year NUM , following a NUM increase in NUM . our eps estimates for NUM and NUM are NUM and NUM , respectively . we have assigned a NUM month stock price target of NUM based on a multiple of NUM times our NUM estimate .

Sample report for semiconductor manufacturer Altera issued by Deutsche Bank on Feb 25th, 2000. Darker shading indicates higher attention weight. The model predicts a market excess return of 0.73% on the next earnings announcement date. Note that the report has been pre-cleaned. Most notably, numbers have been replaced by the "NUM" token.

Figure 3.2: Attention weights

report encoding d_i, I apply two dense transformations to the attentional hidden state,

$$\underset{D_1\times 1}{\tilde{d}_i} = \mathsf{selu}\left(W_{h1}\tilde{h}_i + b_{h1}\right)$$
$$\underset{D_2\times 1}{d_i} = \mathsf{selu}\left(W_{h2}\tilde{d}_i + b_{h2}\right),$$

where selu is the Scaled Exponential Linear Unit activation functions from Klambauer et al. (2017).[4] The choice of the selu activation function here is motivated by two reasons. First, it can be considered a smooth approximation of a rectifier function, $\mathsf{selu}_{s,a}(x) \approx \max(-sa, sx)$, that allows the network to learn piecewise-linear transformation functions. As such, it is closely related to the relu activation function that can be frequently found in neural networks but has some unattractive optimization properties such as being prone to suffer from vanishing or exploding gradients. Second, with the given parameters a and s, selu tends to normalize the input variables which facilitates the training of the network.

[4] The Scaled Exponential Linear Unit is defined as

$$\mathsf{selu}(x) = \begin{cases} s\cdot x & \text{if } x \geq 0 \\ s\cdot a\cdot(\exp(x)-1) & \text{if } x < 0. \end{cases}$$

The parameters $a = 1.67326324$ and $s = 1.05070098$ are chosen such that the output of the selu function has mean zero and unit standard deviation under certain assumption as discussed in Klambauer et al. (2017).

The two-layer specification allows the network to learn more complex non-linear transformation of $\tilde{h}_t$ compared to a network that transforms $\tilde{h}_t$ to d_t with a single layer. In unreported empirical tests, I find that the two-layer specification used here significantly improves the validation loss. This suggests that non-linear transformations play an important role in finding the optimal mapping between analyst reports and announcement returns.

Report aggregator

To obtain announcement-level stock return predictions, the individual report vectors need to be aggregated into a single announcement encoding that represents all the analyst information for the given firm and month. Firm-month encodings are constructed as a weighted average of individual report encodings using an attention mechanism similar to the one used to obtain the report encodings.

To keep things simple, I ignore the sequential nature of analyst reports publications and treat the report encodings for a given firm-month as an unordered collection. The aggregation procedure for reports closely follows the aggregation procedure that was used to aggregate words. I first obtain a hidden representation u_i of the report content by feeding the document vector d_i through a single-layer multi-layer perceptron. This hidden representation is compared to a trainable context vector u_d using the dot product. Similar to the word context vector, the document context vector can be interpreted as the encoding of a fictitious informative report and is learned in the training process. I again transform the alignment score to non-negative weights through with the softmax function. The final firm-announcement vector v is the weighted average of the report encodings multiplied by a scaling factor s^d.

$$\begin{aligned}
\underset{D\times 1}{u_i} &= \tanh(W_d d_i + b_d) \\
\alpha_i &= \frac{\exp(u_i^\top u_d)}{\sum_{i'} \exp(u_{i'}^\top u_d)} \\
s^d &= 1 + \delta_d(M-1) \\
\underset{D\times 1}{v} &= s^d \sum_i \alpha_i d_i
\end{aligned}$$

Together with the attention weights, the scaling parameter δ_d allows the network to learn whether information across different reports is should be summed up or averaged across. For $\delta_d = 0$, the firm-announcement vector v is a simple weighted average of the report vectors d_i with weights α_i that add up to 1. For $\delta > 0$, v can be interpreted as a weighted average of report vectors where the sum of weights $s^d \alpha_i$ is larger than 1.

Note that the report weighting exclusively depends on the content of the reports. An extension of the mechanism could take into account additional information such as the author of the report or the distance between the report publication and the earnings announcement. Such a mechanism could be easily added to the model by introducing a contributor-specific scaling vector that is either added to or multiplied with u_i before calculating the alignment score. In addition, one could take a more direct approach in trying to capture interactions between reports by adding a recurrent layer on top of the report encoder. This approach has the downside that it breaks up the additive structure of the firm encoder, which makes it harder to analyze the market reaction to a report publication with a certain return prediction. I leave the exploration of these ideas for future research.

Predictive regression

Lastly, I use the firm-announcement vector to form predictions of announcement returns. The firm-announcement vector can be thought of as a compressed summary of all return-relevant information that was conveyed by the analysts for a given firm and announcement. This vector is used in a simple linear regression model to form predictions of future returns

$$\hat{r} = W_v v + b_v \tag{3.1}$$

where $\hat{r}$ is a scalar. The affine structure of the report aggregator allows us to express the return prediction $\hat{r}$ as the weighted sum of report-level return predictions $\hat{r}_i$,

$$\hat{r} = \sum_i \alpha_i \underbrace{(W_v d_i + b_v)}_{\hat{r}_i}. \tag{3.2}$$

Eq. (3.2) highlights that the algorithm can only learn interactions between two analyst reports in a fairly limited way. In particular, the algorithm might over- or underweight a particular report based on the presence of another report, since the weight α_i depends on all reports for the given announcement. α_i can be interpreted as a measure of relative signal strength.

The model is solved by minimizing the sum of squared prediction errors over all examples. Using boldface letters to represent vectors of observed and predicted returns for all examples, we can write the optimization problem as

$$\underset{\{\overrightarrow{GRU},\overleftarrow{GRU},W_w,b_w,u_w,W_h,b_h,W_d,b_d,u_d,W_v,b_v\}}{\arg\min} \quad (\hat{\mathbf{r}} - \mathbf{r})^\top \Omega (\hat{\mathbf{r}} - \mathbf{r}) + Regularization \tag{3.3}$$

where $\mathbf{r}$ and $\hat{\mathbf{r}}$ are vectors of the observed and predicted returns, respectively, with their length being equal to the number of examples in the dataset. Ω is a diagonal sample weight matrix and *Regularization* is a penalty term that restricts the weight matrices. Both are explained in detail in

Section 3.2.5.

3.2.4 Pre-training

Pre-trained word embeddings have been proven to significantly improve NLP tasks such as sentence classification (Kim, 2014). Instead of starting the model with a randomly initialized embedding matrix W_e that is then trained together with the remaining neural network, the embeddings are learned separately from the main task, often using a secondary data source. In the general NLP literature, the secondary data source is typically a very large text corpus such as a Wikipedia Dump or a Common Crawl. While this approach has been shown to yield superior embedding matrices that are able to reflect small nuances in word meanings and provide meaningful linguistic representations even of rare words, it cannot be used for the analysis of financial data. The use of current, general-purpose encyclopedias can easily introduce a look-ahead bias, in particular when used for return prediction. To see this, consider the word *"Enron"*. In today's Wikipedia, the word *"Enron"* will be found in close proximity to terms such as *"bankruptcy"* or *"fraud"*, which will impact the word embedding for *"Enron"* beyond its linguistic meaning. Therefore, I pre-train word embeddings using the analyst report corpus itself using *fasttext* (Joulin et al., 2017; Bojanowski et al., 2017) in regular intervals. Fasttext is an extension of Word2Vec (Mikolov et al., 2013a,b) that considers subwords during the training process, i.e. it treats each word as the sum of fixed-length subwords. This is particularly helpful for report texts that were extracted using optical character recognition (OCR), which tend to be noisy. For example, the fasttext algorithm generates an embedding for the word *"earningsat"* - an instance of the phrase *"earnings at"* for which the OCR algorithm failed to recognize the blank space between the two words - that is almost identical to the embeddding for the word *"earnings"* without seeing a large number of instances of *"earningsat"*.

I estimate the word embeddings with a skipgram model and negative sampling. The skipgram model predicts each word token based on its neighboring word tokens using a simple linear transformation of the word embeddings of the neighboring words. The algorithm is summarized in Appendix B.1.1. An in-depth discussion of the skipgram model and negative sampling can be found in Mikolov et al. (2013a,b).

Pre-training the word embeddings has two additional advantages beyond improved prediction performance. First, it guarantees that word embeddings are linguistically meaningful, which improves the interpretability of the results. Second, it allows us to fix the embedding matrix in the training process of the downstream network, which significantly reduces the dimensionality of the training task. The vocabulary size for the full sample is approximately 250,000 after pruning words that occur less than five times. With an embedding size of 128, the resulting embedding matrix has over 30 million free parameters, which compares to 20,611 free parameters of the rest of the model.

Reducing the number of free parameters in the training process reduces the risk of overfitting and significantly speeds up the training process.

3.2.5 Training

Optimization

I estimate the model using stochastic gradient descent with adaptive moment estimation (Adam, Kingma and Ba, 2014) and decaying learning rate on mini-batches of size 64. The model is considered converged if the validation loss does not improve over the course of ten epochs.

Regularization

All matrix transformations outside the GRU layer are restricted by penalizing the l_2-norm of the weight matrices. The regularization term in the loss function Eq. (3.3) can be written as

$$Regularization = \sum_{j=\{w,h1,h2,d,v\}} \lambda_{W_j} W_j + \sum_{j=\{w,d\}} \lambda_{u_j} u_j$$

where the regularization parameters λ are hyperparameters. The GRU layer is not regularized since regularization would restrict its ability to learn long-term dependencies (Pascanu et al., 2013). Instead, I employ dropout after both attention layers (Hinton et al., 2012).

Model instances

Ideally, we would re-train the model every month using the data available at each point in time. However, the training procedure is computationally expensive.[5] To balance computational limitations and the statistical power of a larger and more recent dataset, I re-train the model in coarse intervals. For the first iteration, I use earnings announcements and associated analyst reports up to December 2003. I refer to this model instance as the "2003 model" in the rest of the paper. After that, I retrain the model every two years. For example, for the second iteration, I use announcements up to December 2005, and so on.

Train/validation split and cross-validation

Stock returns exhibit a strong correlation at a given point in time. Therefore, training and validation data must be sampled from distinct time periods. Previous literature (e.g. Gu et al., 2020) suggests splitting the available sample along a single point in time and using the most recent period for validation. I argue that because the time-series correlation of stock returns at monthly frequency is

[5]For example, the 2017 model converged after training for approximately 10h (Nvidia Tesla V100 GPU) plus 4h for pre-training the embeddings (Intel Xeon Platinum 8175, 8 vCPU cores).

fairly weak, a better approach is to randomly select a subset of months from the available sample and use observations from the selected months for validation. This allows me to use more recent data for training. In addition, to the extent that infrequent events such as recessions might dominate stock returns over the period of several months which will be split across the training and validation set in my approach, the randomly sampled validation set might be a better representation of the distribution of returns than a hold-out sample that consists of a number of consecutive months. In particular, I randomly select 10% of the months to form the validation set and use the remaining 90% for training. The validation set is used to fine-tune the hyperparameters and determine the stopping point during the training process.

To maximize the statistical power of the model, I estimate the model using ten-fold cross-validation. Each month of data will be in exactly one validation set of the ten folds, and in the training set of the remaining nine folds. The predictions used in the out-of-sample analysis are the average prediction across the ten folds.

Sample weights

Analyst coverage gradually increased from the beginning of the sample until the mid-2000s. To ensure that the neural network does not overfit the later part of the dataset, I weight examples (firm-announcements) by the inverse of the number of firms in the given quarter in the loss function. The weight is capped at 1/500, which is approximately equal to the 90th percentile of the inverse number of firm-announcements per quarter.

Returns

I use two-day market-adjusted returns to measure the market's reaction to an earnings announcement. The market-adjusted return is calculated as the cumulative stock return from the end of the day before the announcement to the end of the day after the announcement, minus the cumulative return on the value-weighted market portfolio over the same time period. Removing the common market component from the outcome variable reduces the correlation between the examples and makes the algorithm less likely to pick up a common time-series component of returns instead of cross-sectional predictors.

The stochastic optimization procedure is sensitive to large outliers in the data. In addition, large stock price movements are more likely to reflect random noise rather than large expected returns. Keeping large outliers in the dataset is likely to be detrimental to the algorithm's ability to pick up systematic patterns in expected returns. Therefore, I clip the stock returns in the training process at [-0.06,0.06], which is approximately equal to winsorizing at 16% on either side or clipping the variable at ± 1 standard deviations.

Report selection
Each analyst report is assigned to the next earnings announcement following the report publication date. To rule out that the results are diluted by market reactions to the report publication, reports that are published less than three trading days prior to the next earnings announcement are assigned to the subsequent earnings announcement. Reports that are older than 365 days are discarded, as these tend to be assigned to the wrong announcement date due to missing announcement date data. The number of reports assigned to each firm announcement is heavily right-skewed. To reduce the influence of these observations, I cap the number of reports for each announcement at 100. If the dataset contains more than 100 reports for a particular announcement, I rank the reports by contributor and age and discard the lowest ranking reports until the number of reports is no more than 100.[6]

To reduce the size of the dataset and filter out noise, I focus on the first page of each analyst report. Analysts typically summarize their key insights on the first page of the report. While some information on later pages might be relevant for return prediction, these pages are more likely to contain irrelevant text such as generic text blocks or extensive company descriptions. In addition, the substantial reduction of the document size allows for substantially faster training.

Pre-training
Embeddings are pre-trained using a skipgram model with context window size 5. Words that occur less than five times are excluded from the dictionary for computational efficiency. The number of epochs is 5 for a sample of 3 million reports and more, 15 for a sample of 1 million reports and below, and a linear interpolation between [3, 15] for the range of report numbers in between. Bounds are chosen based on the observation that fasttext embeddings tend to under- or overfit outside the [5,15] range and that fewer epochs are required when more i.i.d. samples are available.[7] Embedding vectors are normalized to unit length.

3.2.6 Hyperparameter optimization

The model has 13 hyperparameters, including four parameters that govern the model size and nine regularization parameters. The larger number of parameters and costly training of the model make an exhaustive grid-search approach infeasible. Therefore, I use a Bayesian optimization procedure to choose the hyperparameters as well as the initial learning rate. In particular, I follow Srinivas et al. (2012) and model the hyperparameter choice as a multi-armed bandit problem where the payoff

[6] In other words, the oldest reports of the contributor(s) with the highest number of reports are discarded until there are no more than 100 reports. This approach favors using a more diverse set of reports at the cost of using slightly older reports for stocks with a high analyst coverage.

[7] see e.g. https://fasttext.cc/docs/en/unsupervised-tutorial.html

function is sampled from a Gaussian Process (GP-UCB). The payoff function in this setting is the validation R^2 of the trained model and the inputs are the hyperparameters. At every iteration step, the algorithm estimates an upper confidence bound of the payoff function for all hyperparameter combinations from a grid of the 13 hyperparameters. The algorithm then evaluates the function at the maximum upper confidence bound. In other words, the algorithm trains the model with the most promising set of hyperparameters and records the validation R^2. I run this Bayesian optimization for 60 iterations. 60 iterations are likely to result in a set of parameters that is within close proximity to the optimal parameter set. To set a lower bound on the expected accuracy after 60 iterations, suppose that parameter combinations were sampled at random as in a random search optimization. After 60 trials, the probability that the best randomly sampled parameter combination lies within 5% of the global optimum is $1-(1-0.05)^{60} = 95.4\%$. Of course, the Bayesian optimization does not sample parameter combinations at random. If the payoff function is fairly well described by the Gaussian Process, the expected distance of the obtained parameter combination from the true optimum will be substantially lower.

Due to computational constraints, I optimize the hyperparameters only once using data up to 2003, and use these hyperparameters for the entire sample. The optimal hyperparameters are shown in Table 3.1. The range of explored hyperparameters is shown in Appendix B.1.3. Separate hyperparameter optimization for each model instance would probably improve the performance of the model in later years.[8]

[8]The optimization procedures takes approximately 120h on a Nvidia Tesla V100 GPU using the 2003 sample. The training time scales approximately linearly with the number of examples. In unreported tests, I experiment with using Hyperband (Li et al., 2018) for hyperparameter optimization. Hyperband achieves approximately 80% of the reported validation R^2 of the GP-UCB approach in under 48h.

Parameter	Symbol	Value
Layer dimensions		
Word embedding size	E	128
GRU hidden units (combined directions)	$2H$	16
Report encoder intermediate layer	D_1	256
Report encoder output dimension	D_2	32
Regularization parameters		
Report encoder word attention W_w l_2-regularizer	λ_{W_w}	0
Report encoder word attention u_w l_2-regularizer	λ_{u_w}	10^{-3}
Report encoder W_{h1} l_2-regularizer	$\lambda_{W_{h_1}}$	10^{-6}
Report encoder W_{h2} l_2-regularizer	$\lambda_{W_{h_2}}$	0
Word attention output dropout		0.1
Report attention W_d l_2-regularizer	λ_{W_d}	0
Report attention u_d l_2-regularizer	λ_{u_d}	10^{-8}
Report attention output dropout		0.2
Final prediction W_v l_2-regularizer	λ_{W_v}	10^{-8}

Table 3.1: Neural network hyperparameters

3.2.7 In-sample results

Table 3.2 shows the training statistics. The model predictions explain between 0.68% and 1.20% of the return variation in the training sets and between 0.39% and 0.60% of the return variation in the validation sets. Despite the steeply increasing sample size between the 2003 and 2017 training instances, the validation R^2 does not show a significant time trend. The two rightmost columns report the accuracy for the training and validation set. Accuracy is defined as the fraction of examples for which the neural network predicts the correct sign and can be interpreted as the neural network's ability to predict beats versus misses (in the absence of announcement day risk premiums). In contrast to R^2, accuracy is robust to outliers. Just like the validation R^2, validation accuracy does not exhibit a significant time trend. The absence of a clear time trend in the evaluation metrics suggests that the earlier model instances are not significantly impaired by the smaller sample size.

Validation R^2 and accuracy are uniformly lower than their training counterparts, which suggests that the model has a tendency to overfit the training data. This could potentially be alleviated by further fine-tuning of the regularization parameters.

			R^2 (%)		Accuracy (%)	
Model year	Examples	Reports	Training	Validation	Training	Validation
2003	120,391	654,717	0.68	0.43	52.78	52.24
2005	140,639	938,604	0.81	0.44	52.99	52.41
2007	162,302	1,167,896	0.95	0.51	53.22	52.29
2009	183,307	1,382,554	0.90	0.39	53.20	52.10
2011	204,007	1,631,805	1.02	0.46	53.32	52.35
2013	224,449	1,936,827	0.92	0.45	53.14	52.49
2015	246,351	2,257,973	1.04	0.60	53.23	52.60
2017	268,255	2,578,304	1.20	0.41	53.54	52.13

Examples is the total number of examples (earnings announcements) available to train each model. *Reports* is the number of reports across all examples. Approximately 90% of these examples are used for training, the remainder is used for validation. R^2 is defined in Eq. (3.4). *Accuracy* is the fraction of examples for which the neural network predicts the correct sign.

Table 3.2: Training statistics

3.3 Out-of-sample tests

3.3.1 Out-of-sample predictability

To test the out-of-sample validity of the return predictions, I form predictions for every earnings announcement in the sample starting in the year 2004 using the latest available HAN model instance based on the biennial re-training of the model. In particular, predictions for announcements in the years 2004 to 2005 are based on the 2003 model instance, predictions for the years 2006 to 2007 are based on the 2005 model instance, and so on. Fig. 3.3 plots realized returns against the out-of-sample return predictions. Visual inspection suggests that realized returns are nearly monotonically increasing in the neural return predictions. On average, the slope of a linear regression of realized onto predicted returns is near one. It appears to be slightly lower for positive than for negative returns, suggesting that there is more attenuation in positive predictions than negative predictions. In addition, the binscatter plot suggests that the neural return predictions are unbiased: firms with a positive return prediction tend to have positive announcement day returns, firms with a negative prediction tend to have negative announcement day returns, and firms with a return prediction near zero have an average announcement day return of zero.

To evaluate the out-of-sample performance more formally, I calculate the out-of-sample R^2 as

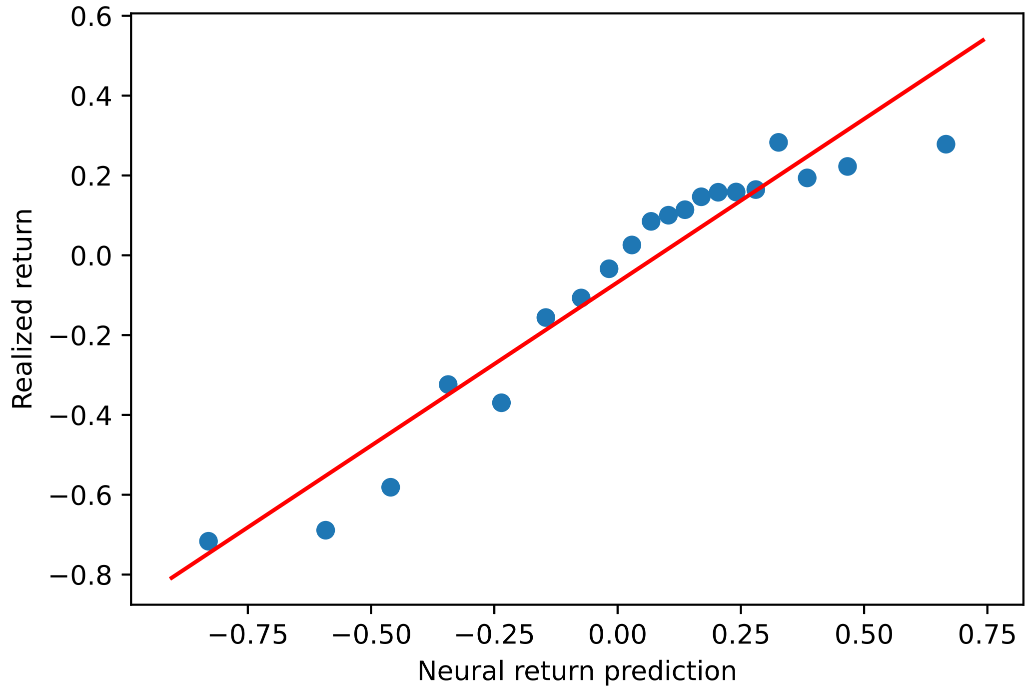
$$R^2_{oos} = 1 - \frac{\sum_{(i,t)} (r_{it} - \hat{r}_{it})^2}{\sum_{(i,t)} r_{it}^2}. \tag{3.4}$$

In addition, I benchmark the neural return predictions against the historical mean,

$$R^2_{oos,mean} = 1 - \frac{\sum_{(i,t)} (r_{it} - \hat{r}_{it})^2}{\sum_{(i,t)} (r_{it} - \bar{r}_t)^2}, \tag{3.5}$$

where the historical mean $\bar{r}_t$ is calculated from the training dataset of the model instance. Ex-ante, it is unclear whether R^2_{oos} or $R^2_{oos,mean}$ impose a tougher benchmark for the model: As Gu et al. (2020) point out, the historical mean is measured with noise, and therefore generally poses a weaker benchmark for out-of-sample return predictions. However, since the unconditional mean return on announcement days is significantly higher than on non-announcement days (Beaver, 1968), this might not be the case on announcement days.

Table 3.3 show the out-of-sample R^2 for the two years after the training period for each model instance. For example, the R^2_{oos} of 0.69% of the 2003 model is the out-of-sample R^2 of the predictions from the 2003 model in the years 2004 and 2005. $R^2_{oos,mean}$ in the same column is the out-of-sample R^2 with $\bar{r}_t$ equal to the average announcement day return r_{it} in the sample ending at



Binscatter plot of out-of-sample realized versus predicted announcement day returns. The neural return predictions are sorted into 20 equally sized bins. The blue dots represent the average predicted return and corresponding realized return of each bin. The red line is a linear regression fit through the original data points. Returns are in percent.

Figure 3.3: Out-of-sample realized vs. predicted announcement returns

the end of the year 2003. The column *All* shows the out-of-sample R^2 over the entire out-of-sample period where the predictions are generated by the latest available model instance at each point in time.

The out-of-sample R^2 for the entire sample is 0.44%. R^2_{oos} for the biennial sub-periods varies between 0.17% and 0.68%. Perhaps unsurprisingly, the highest R^2_{oos} is achieved by the first model instance that was used for hyperparameter training and is therefore likely to be better fine-tuned to the data. R^2_{oos} and $R^2_{oos,mean}$ are very similar for the full sample. The results suggest that the neural network is able to learn persistent predictors of announcement returns. R^2_{oos} is positive in every two-year subperiod and statistically significant in seven out of eight subperiods, including the financial crisis.

The out-of-sample R^2 for the entire sample of 0.44% is remarkably similar to the average validation R^2 of 0.46% shown in Table 3.2. This confirms that the randomly sampled validation sets are sufficiently independent of the training data to serve as validation samples for hyperparameter tuning and early stopping.

Model OOS period	2003 04-05	2005 06-07	2007 08-09	2009 10-11	2011 12-13	2013 14-15	2015 16-17	2017 18-19	latest 04-19
R^2_{oos}	**0.68**	**0.43**	**0.36**	**0.36**	**0.62**	**0.46**	0.17	**0.49**	**0.44**
	(5.01)	(3.81)	(3.45)	(3.17)	(5.52)	(3.98)	(1.42)	(3.99)	(10.57)
$R^2_{oos,mean}$	**0.69**	**0.47**	**0.36**	**0.38**	**0.63**	**0.46**	0.16	**0.49**	**0.44**
	(5.14)	(4.18)	(3.41)	(3.31)	(5.56)	(3.96)	(1.38)	(3.98)	(10.75)

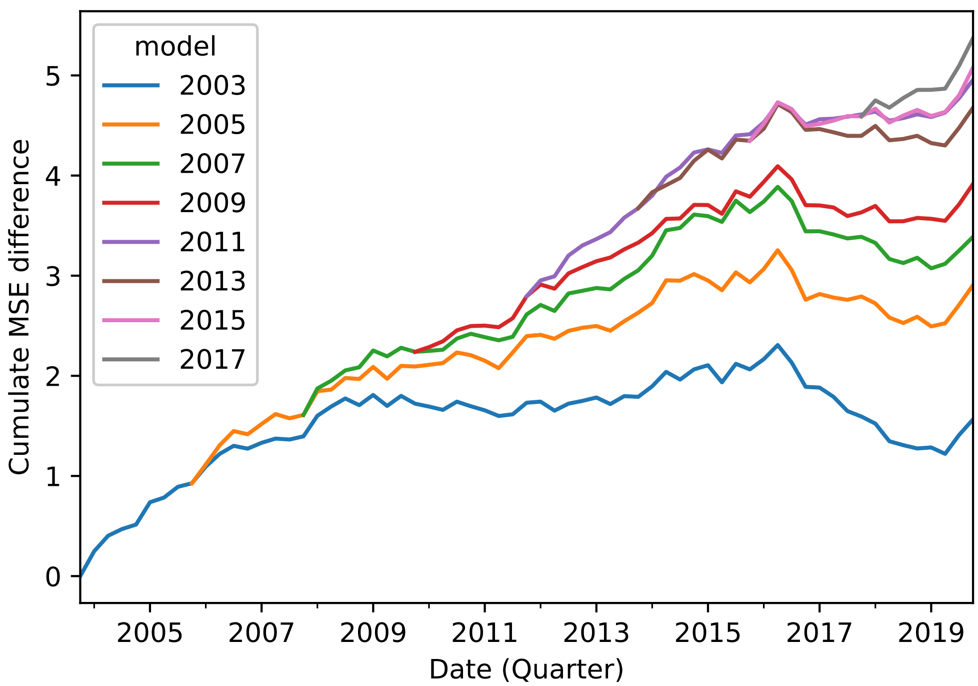
The table shows out-of-sample R^2 measures for the two years after the model formation. R^2_{oos} is the out-of-sample R^2 with benchmark zero as shown in Eq. (3.4). $R^2_{oos,mean}$ is the out-of-sample R^2 with the historical mean benchmark as shown in Eq. (3.5). Diebold-Mariano test statistics are shown in parentheses. Boldface indicates statistical significance at the 1% level.

Table 3.3: Out-of-sample R^2

3.3.2 On the persistence of predictors

If the predictive patterns present arbitrage opportunities to investors, we expect them to disappear over time. To study the persistence of the predictors that were identified by the neural network, I follow Welch and Goyal (2008) and plot the difference in cumulative quarterly mean squared errors between the historical mean model and the neural return predictions in Fig. 3.4. An increasing line indicates that the neural return predictions outperform the historical mean model. I plot the mean squared error rather than the sum of squared errors to account for the varying number of observations across quarters, effectively giving equal weight to each quarter in the out-of-sample period. The visual analysis suggests that the predictive power of each model lasts for several years but slowly dissipates. At the end of the sample period, the cumulative MSEs of the eight models line up almost perfectly with the model years. With the exception of the 2013 model which has a lower cumulative performance than the 2011 model, newer models always have a higher cumulative performance than all of the older models.

The 2003 model continues to predict returns for about half a decade and then flattens out significantly. This suggests that the return predictors identified by the 2003 model hold for an extended period of time but eventually vanish. We can observe a similar, albeit less pronounced, flattening of the cumulative MSE curves of the other models over time. I test the hypothesis that the predictive power diminishes over time more formally by testing the hypothesis that the neural network outperforms the historical mean model in every two-year period following the training year by conducting moving-window Diebold-Mariano tests. Results are shown in Table 3.4. The tests confirm that each model instance is able to produce positive and significant out-of-sample R^2 for four to eight years, with the exception of the 2016-2017 period. The Diebold-Mariano test statistic is highest in the out-of-sample period that immediately follows the model training for all model instances. The findings in this section suggest that investors eventually learn from past mistakes when it comes to forming optimal earnings expectations based on the textual information in the analyst reports, but only do so very slowly over the course of several years.



The figure shows the out-of-sample performance of the neural return predictions. In particular, it shows the cumulative quarterly difference in the mean squared prediction error of the neural network versus the mean squared prediction error of the prevailing historical mean. For better visualization, each model instance starts with the cumulative mean difference of the previous model, e.g. the 2005 model starts with the cumulative MSE difference of the 2003 model in 2005Q4. The first model instance, 2003, starts with a value of 0 in 2003Q4.

Figure 3.4: Quarterly performance of different model instances

Model OOS period	2003	2005	2007	2009	2011	2013	2015	2017
04-05	**5.14**							
06-07	**2.73**	**4.18**						
08-09	1.79	**2.68**	**3.41**					
10-11	0.04	1.67	2.10	**3.31**				
12-13	0.28	1.23	**2.39**	**3.22**	**5.56**			
14-15	1.39	1.52	**3.12**	**2.74**	**4.83**	**3.96**		
16-17	-2.32	-0.68	-1.26	-0.88	1.32	0.32	1.38	
18-19	-0.13	0.58	0.03	1.66	**2.35**	1.70	**2.84**	**3.98**

The table shows Diebold-Mariano test statistics for the out-of-sample return predictions benchmarked against the historical mean model. Columns indicate the model instance and rows indicate the out-of-sample period. Positive DM values indicate that the $R^2_{oos,mean}$ is larger than zero. Boldface indicates statistical significance at the 1% level of a one-tailed test.

Table 3.4: Diebold-Mariano test statistics for further-ahead out-of-sample periods

3.4 Information diffusion

One of the most studied empirical artifacts in financial markets is the delayed price response to publicly available news. A myriad of papers has documented that realized earnings surprises are not fully reflected in prices immediately, but drift towards the new equilibrium price over the course of several trading days, known as the post-earnings announcement drift (e.g. Ball and Brown, 1968; Bernard and Thomas, 1989). The delay is stronger when investor attention is lower and processing costs are higher (Hirshleifer et al., 2009). Investor attention to analyst reports is likely to be lower than to earnings announcements, and processing costs of textual information are almost certainly higher than those for numerical information, so we would expect to see similarly delayed price responses for analyst reports. To test whether the observed predictability can be explained by gradual diffusion of the information in the reports, I run an event study on report publication and earnings announcement days conditional on the neural return predictions. At this point, the affine structure of the report aggregator demonstrated in Eq. (3.2) comes in handy: while the neural network is designed to make a single prediction based on a number of reports, we can represent this prediction by a weighted average of predictions that were made based on a single report. This allows me to study the market reaction to a single report. Recall that the report predictions were formed to predict returns on the next earnings announcement day that is at least two trading days after the report publication day. To emphasize the difference between the market reaction to the release of the reports versus the release of the actual earnings, I refer to returns on report publication days as *publication returns* and returns on earnings announcement days as *announcement returns*.

The left panel in Fig. 3.5 shows the abnormal returns around the report publication conditional on the report-level return prediction $\hat{r}_i$ as shown in Eq. (3.2). The plot reveals that the publication returns are positively correlated with the announcement prediction. Report publications with a neural return prediction (NRP_i) in the highest quintile coincide with an abnormal publication return of around 1%, and publications with a neural return prediction in the lowest quintile coincide with a return of around -0.5%. The figure also shows significant drifts leading up to the publication of both negative and positive NRP_i reports. This suggests that the neural return prediction does not only capture information that was revealed by the report itself but also captures information that was revealed to at least a subset of investors beforehand and then reiterated in the analyst report. An example of public information that is frequently reiterated in analyst reports is an earnings call. Analysts often publish new reports after earnings calls and summarize their takeaways from the earnings call in these reports.

The post-publication drift appears to be very small. Returns after the publication of low NRP reports are virtually flat, and only marginally positive after the publication of high NRP reports. This suggests that the neural network does not simply pick up post-publication drifts of reports that are published close to the earnings announcement.

The right panel in Fig. 3.5 shows the abnormal return around earnings announcements conditional on the *firm-level* return prediction $\hat{r}$. The figure confirms the predictive power of the neural network. Firms in the highest NRP quintile have a positive abnormal return on the announcement day and firms in the lowest NRP quintile have a negative abnormal return on the announcement day. Together with the findings from the publication days, this suggests that the neural network on average predicts underreaction to both negative and positive news. In contrast to the publication reactions, the magnitude of the abnormal return is higher for negative predictions than for positive predictions, which suggests that the underreaction to negative news on the publication date is stronger than the underreaction to positive news.

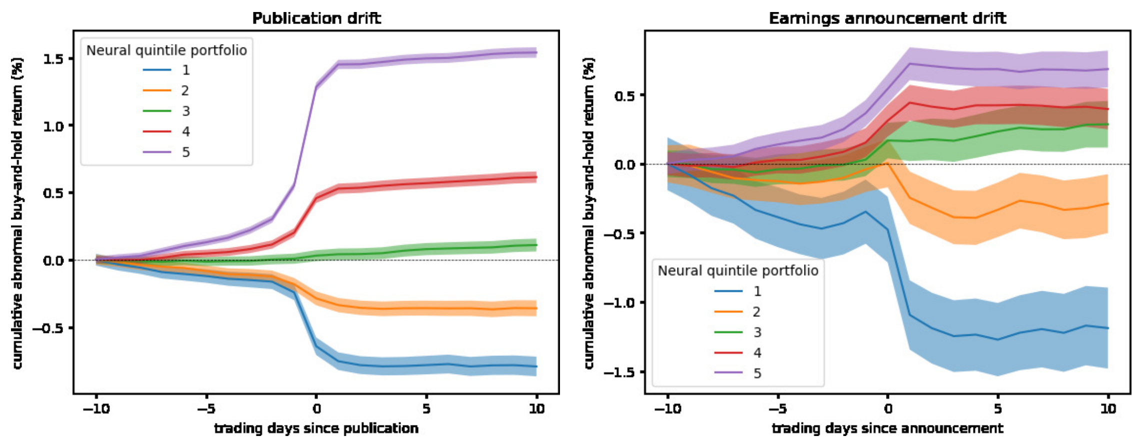
In contrast to typical earnings announcement event studies, the event study here has two information events that are unequally spaced and potentially overlapping. To tease apart the effects of publications and announcements, I run a joint event study by estimating a single regression of daily returns on dummy variables that indicate the distance to both the publication as well as the announcement. In particular, I estimate

$$\begin{aligned} r_{it} = \beta_t &+ \sum_{\tau \in \mathrm{T}} \beta_{t\tau}^{pub} NRP_{it}^{report} 1(\text{pub+}\tau \text{ days})_{it}^{report} \\ &+ \sum_{\tau \in \mathrm{T}} \beta_{t\tau}^{act} NRP_{it}^{report} 1(\text{actual+}\tau \text{ days})_{it}^{report} + \varepsilon_{it}^{report} \end{aligned} \tag{3.6}$$

in a two-stage Fama-Macbeth regression. The unit of observation is a report-date, i.e. the sample contains all reports and the associated returns in a [-30,70] trading day window around the report publication. To be clear, a single r_{it} can appear in the sample multiple times if the [-30,70] trading day window of one report overlaps with the window of another report of the same firm. The overlapping windows and repeated sampling of r_{it} induce cross-sectional correlation of the residuals in the first stage, but not in the second stage. 1(pub+τ days) is an indicator variable that is equal to one if the return is τ days after the publication of the report. Similarly, 1(act+τ days) is equal to one if the return is τ days after the earnings announcement date. For both publications and announcements, I look at a ±5 day window around the event and capture the days outside the event window in a single dummy variable on either side of the window, i.e. $\mathrm{T} = \{\text{less than } -5, -5, -4, \ldots, 4, 5, \text{more than } 5\}$. Results of the second stage are shown in Table 3.5. A NRP_i of 1% coincides with a 0.605% return on the publication date. An increase of NRP_i by the same magnitude predicts a 0.630% higher return on the day after the earnings announcement. Since most earnings are announced during after-hours, *actual +1 day* marks the first trading day after the earnings announcement in most cases. The coefficient on the announcement day dummy is very close to the coefficient on the publication day dummy, suggesting that the neural network captures information that is on average only halfway incorporated into prices upon publication. The regression results show that prices do not exhibit any statistically significant drift around either the publication or the announcement day when the drifts are estimated jointly. The apparent drifts in Fig. 3.5 are most likely artifacts of overlapping event windows of publications and announcements.

The lack of a statistically significant drift between publication and announcement suggests that the predictability of announcement returns is distinct from slow information diffusion. Instead, the information in the report does not seem to diffuse at all beyond the initial reaction. (Cohen et al., 2020) observe a similar pattern in the market reactions to changes in the risk section of 10-K filings: while changes of the text in the risk section have strong predictive power for future earnings and returns, there is no price drift after the initial publication, and markets only react once the described event is realized.

The lack of a post-announcement drift shows that the information that was picked up by the HAN is incorporated into prices quickly upon the announcement of the actual earnings. Once the earnings are announced, investors seem to respond to the “news“ quickly, suggesting that attention to the news is likely to be high and limits to arbitrage low on the actual earnings announcement day.



Cumulative abnormal returns (Fama and French, 2015, plus Momentum) around report publications and subsequent earnings announcements. Publications are on report level but announcements are on firm level. Shaded regions are 95% confidence bands.

Figure 3.5: Publication and announcement drifts

	Coefficient	t-stat
Intercept	0.006*	1.688
$NRP \times 1$(more than 5 days before pub)	-0.090	-1.007
$NRP \times 1$(pub -5 days)	-0.079	-0.863
$NRP \times 1$(pub -4 days)	-0.091	-1.013
$NRP \times 1$(pub -3 days)	-0.097	-1.070
$NRP \times 1$(pub -2 days)	-0.086	-0.931
$NRP \times 1$(pub -1 day)	0.112	1.026
$NRP \times 1$(pub)	0.605***	6.347
$NRP \times 1$(pub +1 day)	0.116	1.227
$NRP \times 1$(pub +2 days)	-0.053	-0.555
$NRP \times 1$(pub +3 days)	-0.112	-1.231
$NRP \times 1$(pub +4 days)	-0.109	-1.182
$NRP \times 1$(pub +5 days)	-0.110	-1.173
$NRP \times 1$(more than 5 days after pub)	-0.123	-1.388
$NRP \times 1$(more than 5 days before actual)	0.126	1.383
$NRP \times 1$(actual -5 days)	0.139	1.324
$NRP \times 1$(actual -4 days)	0.257*	1.720
$NRP \times 1$(actual -3 days)	-0.105	-0.366
$NRP \times 1$(actual -2 days)	-0.154	-0.535
$NRP \times 1$(actual -1 day)	0.198	1.512
$NRP \times 1$(actual)	0.192	0.984
$NRP \times 1$(actual +1 day)	0.630***	3.315
$NRP \times 1$(actual +2 days)	-0.022	-0.133
$NRP \times 1$(actual +3 days)	0.123	0.903
$NRP \times 1$(actual +4 days)	0.033	0.285
$NRP \times 1$(actual +5 days)	0.100	0.880
$NRP \times 1$(more than 5 days after actual)	0.117	1.289

The table shows the second stage of a Fama-Macbeth regression of daily abnormal returns (Fama and French, 2015, plus Momentum) on the neural return prediction around the report publication date and subsequent earnings announcements as shown in Eq. (3.6). The unit of observation is a report in the first stage and a date in the second stage. 1(pub) is a dummy variable that is equal to one on publication days, 1(pub -1 day) is a dummy variable that is equal to one one trading day before the publication, 1(more than 5 days before pub) is equal to one for all trading days that are more than 5 days prior to the publication, etc. 1(actual) is equal to one on earnings announcement days. The remaining indicator variables are defined accordingly. *NRP* is the out-of-sample neural return prediction of the most recent model at the time of report publication. The sample includes all trading days within a [-30,70] trading day window around each report publication from 2004 to 2019.

Table 3.5: Publication and announcement drifts

The evidence presented up to this point suggests that markets do immediately not incorporate the information picked up by the neural network into prices upon publication. This opens the question whether the neural network identifies very particular types of information that lead to the observed predictability, or whether markets underreact to information in analyst reports more broadly and the neural network simply picks up this general pattern. In the latter case, the average publication reaction to analyst reports preceding the earnings announcement would be a sufficient statistic of the predictive information in the neural return predictions. To test this hypothesis, I regress announcement reactions onto the out-of-sample return predictions controlling for the average publication reaction. The average announcement reaction is the average of the abnormal publication returns of the reports that were used to form the neural return predictions. The abnormal publication return is the cumulative abnormal return realized in a three-day window around the report publication. Abnormal returns are measured relative to a Fama-French four-factor model, and the three-day window starts with the closing price on the day prior to the publication day and ends with the closing price on the second trading day after the publication. To establish a benchmark, I first run a univariate regression of realized returns on predictions. Results are shown in Table 3.6 column (1). The OLS coefficient on the neural return prediction is approximately 0.84. A coefficient below one in the out-of-sample test suggests that there is some attenuation of the in-sample predictability, which is in line with the observed difference between in-sample and out-of-sample R^2 in Section 3.2.7. Controlling for market-wide factors has little impact on the predictive power of the neural network (column (2)). In column (3), I replace the neural return prediction with the average return on the day of the analyst report publication on the right-hand side of the regression. The coefficient estimate of approximately 0.01 is statistically significant at the 5% level but economically small: a publication return of 1% predicts an announcement return of 1bps, and the within-R^2 for publication returns is zero. In column (4), I regress the announcement return on both the neural return prediction and the average announcement return. Controlling for the publication return has virtually no impact on the predictive power of the neural return prediction. The coefficient estimate shrinks marginally from 0.725 to 0.723, and the R^2 remains unchanged up to the second significant digit. However, controlling for NRP renders the coefficient on the publication return insignificant.

The findings suggest that the persistent underreaction to information in analyst reports is not a general feature of market reactions to analyst reports. Instead, the neural network seems to pick up particular information that is only partially incorporated into prices upon publication. Put differently, the finding rejects the hypothesis that the predictive power of the neural network can be explained by slow reaction to news to the extent that prices react slowly at a constant rate as suggested by Daniel et al. (2020), among others.

	(1)	(2)	(3)	(4)
Neural return prediction	0.841***	0.725***		0.723***
	(18.28)	(15.67)		(15.61)
Publ. return			0.009**	0.002
			(2.35)	(0.61)
Constant	-0.068***			
	(-3.37)			
Fixed Effects		(t)	(t)	(t)
Observations	170,028	169,983	169,983	169,983
R^2	0.0048	0.0401	0.0371	0.0401
Within R^2	0.0048	0.0031	0.0000	0.0031

The table shows regressions of announcement returns on out-of-sample predictions and publication returns. The dependent variable is the market-adjusted return in the two-day announcement window. Columns (2)-(4) use daily date fixed effects as indicated in the fixed effects row. Since every announcement reaction is measured over the span of two trading days and the two-day windows of different firms might only partially overlap, every observation has two date-fixed effects corresponding to the two days in the return measurement window. Standard errors are triple-clustered by firm and the two dates. t-Statistics are shown in parentheses.

Table 3.6: Neural return prediction vs. publication returns

3.5 ... and the cross-section of returns

3.5.1 Portfolio sorts

In the previous two sections, I showed that the HAN can predict returns on earnings announcement dates. A number of papers have suggested that earnings announcement days play an important role in explaining asset pricing anomalies. For example, Engelberg et al. (2018) documents that anomaly returns are significantly higher on earnings announcement days than on non-earnings announcement days, and Lochstoer and Tetlock (2020) argue that anomaly returns are dominated by cash flow news, which tends to be highest on earnings announcement days. To investigate the link between the predictability of earnings announcement returns and the broader literature on cross-sectional predictability of stock returns, I form monthly portfolios of stocks based on the return predictions of the neural network. Every month, I use the latest available HAN model instance based on the biennial re-training of the model (see Section 3.2.5) to predict the next earnings announcement return. In contrast to the previous section, I do not restrict myself to firms with earnings announcements here. Instead, for every firm and month, I form predictions based on the reports published between the last earnings announcement and the end of the current month. As before, to rule out that my analysis captured the contemporaneous market reaction to the news revealed in the report, I discard reports published less than three trading days before the end of the month.

At the end of every month, I sort firms into five equally-sized portfolios based on their NRP for the next earnings announcement date and hold them for one month. Results are shown in Table 3.7. An equally-weighted long-short portfolio earns a highly significant CAPM alpha of 93bps per month and an abnormal six-factor (Fama and French, 2015, plus momentum) return of 49bps per month. Abnormal returns increase monotonically with the neural return prediction.

Despite not having supplied any information about firm characteristics to the HAN, the factor betas reveal that the neural network picked up several well-known anomalies that were discovered using characteristics-based portfolios. The neural long-short portfolio loads negatively on the market (Frazzini and Pedersen, 2014) in the CAPM regression, and positively on momentum (Jegadeesh and Titman, 1993) and profitability (Novy-Marx, 2013a) in the six-factor regression. Surprisingly, the neural long-short portfolio loads negatively on size (Fama and French, 1992). However, it is worth noting that the return on the size factor was economically and statistically insignificant at just 2bps per month during the out-of-sample period and negative for most of the model training period of the 1980s and 1990s.

3.5.2 Time series

The relatively short sample period of 192 months makes it difficult to further split the sample into subperiods. To investigate whether the predictive power of the NRP is generated in only a subset of the sample, I plot the cumulative gross return and the cumulative abnormal in Fig. 3.6. Visual inspection suggests that the abnormal returns occur throughout the entire sample with a notable reversal coming out of the 2007-2009 recession.

3.5.3 A closer look at the profitability anomaly

Section 3.5.1 revealed a strong relationship between the profitability anomaly and the neural return predictions. In this section, I test whether the NRP portfolio spans the profitability-sorted portfolio. To do so, I use the high minus low predictability announcement return prediction portfolio, which I refer to as *NRP* in the following, as an explanatory variable in various factor regressions with the profitability factor *RMW* on the left-hand side. Results are shown in Table 3.8. The NRP portfolio captures around two-thirds of the alpha on the RMW portfolio in the 2004-2019 period. The remaining CAPM alpha is only marginally significant, while the p-values on the remaining FF4

	Low	2	3	4	High	High-Low
Intercept	-0.795***	-0.258	-0.159	0.011	0.138	0.933***
	(-2.98)	(-1.44)	(-1.10)	(0.09)	(1.30)	(3.95)
Mkt-RF	1.494***	1.371***	1.288***	1.214***	1.173***	-0.321***
	(22.80)	(31.24)	(36.59)	(39.26)	(44.96)	(-5.55)
N	192	192	192	192	192	192
$Adj.R^2$	0.73	0.84	0.88	0.89	0.91	0.14

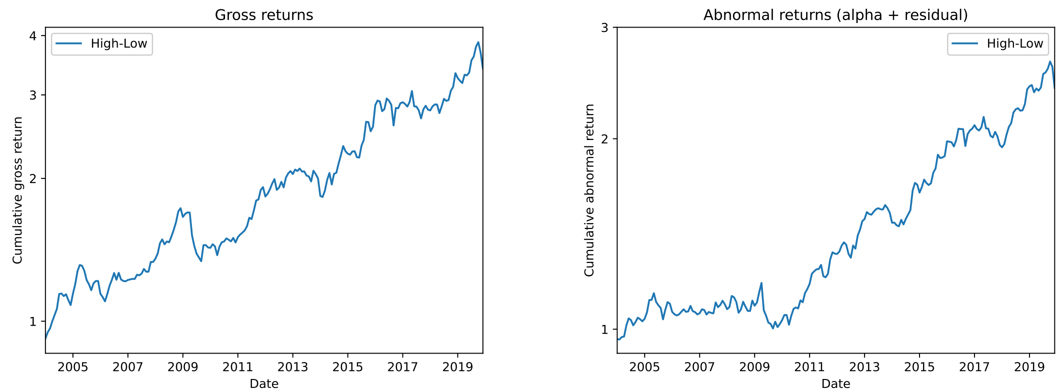
Panel A: CAPM

	Low	2	3	4	High	High-Low
Intercept	-0.295*	0.034	0.022	0.116**	0.192***	0.487***
	(-1.85)	(0.40)	(0.38)	(2.09)	(2.97)	(2.61)
Mkt-RF	1.092***	1.054***	1.049***	1.021***	1.037***	-0.055
	(23.83)	(42.50)	(63.09)	(64.31)	(55.72)	(-1.03)
SMB	1.086***	0.856***	0.693***	0.675***	0.563***	-0.524***
	(14.54)	(21.16)	(25.55)	(26.07)	(18.54)	(-5.99)
HML	-0.231***	0.239***	0.323***	0.238***	0.031	0.262***
	(-2.97)	(5.67)	(11.42)	(8.82)	(0.98)	(2.87)
CMA	-0.138	-0.098	-0.041	-0.101**	-0.171***	-0.033
	(-1.10)	(-1.45)	(-0.91)	(-2.33)	(-3.37)	(-0.23)
RMW	-0.930***	-0.223***	0.009	0.118***	0.091**	1.021***
	(-8.59)	(-3.81)	(0.22)	(3.15)	(2.08)	(8.06)
Mom	-0.156***	-0.177***	-0.129***	-0.090***	-0.066***	0.090*
	(-3.93)	(-8.28)	(-9.00)	(-6.58)	(-4.10)	(1.93)
N	192	192	192	192	192	192
$Adj.R^2$	0.91	0.96	0.98	0.98	0.97	0.50

Panel B: FF6

The table shows regression results for equally-weighted portfolios sorted on the neural return prediction. At the end of each month, firms are sorted into quintiles based on their neural return prediction. *Low* denotes the lowest predicted return quintile, and *high* denotes the highest predicted return quintile. The sample period is January 2004 to December 2019. Returns are monthly and in percent. T-statistics are shown in parentheses.

Table 3.7: Portfolios sorted on neural return prediction



The figure shows the cumulative returns of a long-short portfolio based on neural return predictions. The long-short portfolio is the monthly rebalanced portfolio that goes long firms in the highest quintile of neural return predictions and short firms in the lowest quintile of neural return predictions. The left panel shows the cumulative gross return and the right panel shows the cumulative abnormal return (alpha + residual from a FF6 regression).

Figure 3.6: Cumulative returns of the neural long-short portfolio

and FF4+CMA alpha exceed 10%.

	RMW	RMW	RMW	RMW	RMW
Intercept	0.396***	0.375***	0.165*	0.152	0.153
	(3.70)	(3.59)	(1.76)	(1.63)	(1.63)
Mkt-RF	-0.160***	-0.122***	-0.080***	-0.075***	-0.076***
	(-6.09)	(-4.09)	(-3.37)	(-2.93)	(-2.90)
SMB		-0.189***		-0.007	-0.006
		(-3.88)		(-0.15)	(-0.14)
HML		-0.024		-0.089**	-0.085*
		(-0.46)		(-2.26)	(-1.84)
CMA		-0.031			-0.015
		(-0.37)			(-0.20)
Mom		-0.012		-0.032	-0.032
		(-0.46)		(-1.39)	(-1.37)
NRP			0.248***	0.255***	0.254***
			(8.98)	(8.09)	(8.06)
N	192	192	192	192	192
$Adj.R^2$	0.16	0.21	0.41	0.42	0.41

The table shows regressions of the profitability factor (RMW) onto other factors as well as the neural return prediction portfolio (NRP). Mkt-RF, SMB, HML, CMA are the factors from the Fama-French five-factor model, Mom the 12-2 Momentum factor, and NRP a portfolio that goes long high neural return predictions and short low neural return predictions.

Table 3.8: Using NRP to explain the profitability anomaly

3.6 Can risk premiums explain the predictability of announcement returns?

3.6.1 Market risk

To rule out a risk-based explanation for the observed predictability of announcement day returns, I run several tests in the spirit of Savor and Wilson (2016). First, I test whether the neural return predictions predict higher market betas on the announcement day. Results are shown in Table 3.9. I find no support for the hypothesis that market beta explains the return predictability. The coefficient estimate on the interaction term between the neural return prediction and the market return is marginally negative and insignificant.

	Return
Neural return prediction	0.552***
	(14.15)
Mkt-RF	0.997***
	(27.65)
Neural return prediction × Mkt-RF	-0.012
	(-0.43)
Constant	0.028
	(0.84)
Observations	162,984
R^2	0.058

The table shows regressions of daily announcement day returns on the market return and the neural return predictions. Standard errors clustered by firm and month are shown in parenthesis.

Table 3.9: Announcement day betas

3.6.2 Aggregate growth

Next, I test whether the cash flow news channel proposed by Savor and Wilson (2016) can explain the observed announcement returns. In their model, announcing firms carry a risk premium because their earnings announcements are informative about aggregate cash flows. In the cross-section, firms that are positively correlated to future aggregate earnings growth should carry a positive risk premium, and firms that are negatively correlated to future aggregate earnings growth should carry a negative risk premium. I test this hypothesis by regressing future aggregate earnings growth on the return on the high and low neural return prediction portfolio. Results are shown in Table 3.10. Again, I do not find support for the hypothesis that risk explains the return predictability: neither the high nor the low return NRP portfolio significantly predict future earnings growth. Moreover, the coefficient estimate on the low portfolio is positive, while the strong predictability of negative announcement returns would require the portfolio to have a strong negative correlation with future earnings growth.

	Earn. Growth$_t$	Earn. Growth$_t$
Low NRP$_{t-1}$	0.436	
	(1.36)	
High NRP$_{t-1}$		0.534
		(1.26)
Constant	-0.006	-0.017
	(-0.25)	(-0.60)
Observations	64	64
R^2	0.042	0.032

The table shows regressions of quarterly earnings growth on the returns of the high and low neural return prediction portfolio in the previous quarter. T-statistics based on Newey-West standard errors with four lags are shown in parenthesis.

Table 3.10: Predicting future earnings growth with NRP portfolios

3.6.3 Discount rate news

An alternative explanation for the observed announcement reaction would be that the neural network predicts discount rate news. To test this hypothesis, I estimate each firm's loadings on the five factors from Fama and French (2015) and the Momentum factors both before and after the announcement. In particular, I estimate the pre-announcement loadings in a 90 calendar day window ending 10 days prior to the announcement, and the post-announcement loadings in a 90 calendar day window starting 10 days after the announcement. The gap between the two windows ensures that neither of the two estimates is impacted by the announcement day reaction. I then run regressions in the form

$$\Delta\beta_{F,it} = \beta_{F,it}^{post} - \beta_{F,it}^{pre} = \delta_0 + \delta_1 NRP_{it} + \varepsilon_{it} \quad (3.7)$$

where F is one of the six factors. Results are shown in Table 3.11. I find that *NRP* does not predict the change in the loading on any of the six factors. This lends further support to the hypothesis that *NRP* predicts cash flow news.

	$\Delta\beta_{Mkt}$	$\Delta\beta_{SMB}$	$\Delta\beta_{HML}$	$\Delta\beta_{MOM}$	$\Delta\beta_{CMA}$	$\Delta\beta_{RMW}$
NRP	-0.002	-0.003	0.000	0.007	-0.044	0.067
	(-0.15)	(-0.12)	(0.00)	(0.28)	(-0.79)	(1.54)
Constant	-0.000	-0.001	-0.004	-0.009	0.003	-0.010
	(-0.05)	(-0.23)	(-0.45)	(-1.13)	(0.21)	(-0.90)
N	126,929	126,929	126,929	126,929	126,929	126,929
R^2	0.00	0.00	0.00	0.00	0.00	0.00

The table shows regressions of changes in six-factor betas around the earnings announcement on the neural return prediction (NRP). $\Delta\beta_F$ is the difference between the post-announcement loading and the pre-announcement loading on factor F. Pre (post) announcement betas are measured in a 90 calendar day window ending (starting) 10 days prior to (after) the earnings announcement date. Each firm has at least 45 valid returns in both the pre and post announcement window. Standard errors are clustered by time (month). t-Statistics are shown in parentheses.

Table 3.11: Change in factor betas

3.6.4 Analyst expectations

The previous tests suggest that neural return predictions capture biased cash flow expectations rather than risk premiums. If the aggregate market expectations are biased, individual market participants' expectations are likely to be biased in the same direction. To find further support for this hypothesis, I investigate whether NRP predicts biases in the stated expectations of one particular group of market participants: stock market analysts. To do so, I calculate realized earnings surprises as well as revisions in longer-term earnings expectations around earnings announcement dates with respect to the median analyst forecast. Earnings surprises are defined as

$$Surp_t = \frac{E_t - \mathbb{E}_{t-1}[E_t]}{P_{t-45}}. \tag{3.8}$$

where E_t are the quarterly earnings per share announced at t. $\mathbb{E}_{t-1}[E_t]$ is the consensus forecast of E_t measured at $t-1$, the day before the announcement. P_{t-45} the stock price 45 days before the announcement. Realized earnings tend to be noisy and dominated by transitory earnings shocks with limited impact on a company's stock price. Therefore, I also calculate revisions in the analysts' longer-term earnings expectations,

$$Rev_{t,t+\tau} = \frac{\mathbb{E}_t[E_{t+\tau}] - \mathbb{E}_{t-1}[E_{t+\tau}]}{P_{t-45}}. \tag{3.9}$$

Here, $\mathbb{E}_t[E_{t+\tau}]$ is the consensus forecast of future earnings in $t+\tau$ measured right after the announcement of current earnings at t. In Table 3.12, I show that NRP strongly predicts both earnings surprises and revisions of longer-term earnings expectations at the earnings announcement date. This suggests that both short-term and long-term analyst forecasts are biased in the same direction as the HAN predicts earnings announcement returns.

	Surp_t	$\text{Rev}_{t,t+1Q}$	$\text{Rev}_{t,t+1Y}$	$\text{Rev}_{t,t+2Y}$
Neural return prediction	0.780***	0.468***	1.617***	1.287***
	(8.85)	(12.67)	(12.16)	(13.36)
Constant	-0.230***	-0.267***	-0.666***	-0.518***
	(-7.54)	(-16.68)	(-13.01)	(-12.10)
Observations	140,876	136,728	130,807	110,799
Adj. R^2	0.0045	0.0100	0.0103	0.0124

The table shows regressions of consensus forecast errors and consensus forecast revisions on the neural return prediction. $Surp_t$ is the realized quarterly EPS at time t minus the consensus forecast prior to the earnings announcement, where t is the earnings announcement date. $Rev_{t,t+\tau}$ is the revision of the τ-periods ahead consensus EPS estimate at time t. Pre- and post-earnings consensus forecasts are measured within a 45-day window before and after the announcement, respectively. $Surp$ and Rev are scaled by the stock price 45 days prior to the announcement. The consensus forecast is the mean forecast published in the window. If an analyst publishes multiple forecasts within the window, only their latest forecast is considered. Standard errors are double clustered by firm and time (month). t-Statistics are shown in parentheses.

Table 3.12: Analyst surprises and revisions

3.7 Model diagnostics

3.7.1 Explaining the neural network with sparse local approximations

Neural networks are powerful because of their ability to model complex transformations and interactions between the input features. However, these interactions make the predictions very hard to interpret by a human observer, which is why deep learning algorithms are often labeled as "black box" methods. The lack of interpretability is among the main reasons for the rather slow adoption of machine learning techniques in economics, where causal inference is often more important than predictive power alone.

Several methods to interpret machine learning methods have emerged from the literature. Baehrens et al. (2010) suggest interpreting the gradient vector as a measure of attribution of input features to the model prediction. This approach was adopted in the finance literature by Chen et al. (2020). However, the partial derivatives in the gradient vector are an imperfect measure of feature importance in a model with strong feature interactions. In an instance where a combination of features is important for a particular prediction, the gradients do not represent a meaningful allocation of the joint importance to the individual features. Moreover, gradients might be misleading in particular for very confident predictions where the gradient is near zero. An alternative approach to measuring feature importance is to train a surrogate model that explains the predictions of the machine learning model. Ribeiro et al. (2016) take this approach and study predictions of machine learning models through sparse, locally linear approximations of the model. While the global decision function of a neural network is usually highly non-linear, the function can possibly be approximated by a linear function in a small neighborhood around the actual example. We can obtain such locally linear approximations by applying small perturbations to the observed input data, i.e. removing or replacing individual words within a report, and fitting a linear model to the perturbed data. If the local linearity assumption is valid, we can attribute the observed change in the neural network's prediction across the perturbations to the individual words.

I follow Ribeiro et al. (2016) and obtain word attributions for a randomly selected subset of reports following the LIME procedure (local interpretable model-agnostic explanations). For each document, explanations are obtained as follows:

1. Perturb the document 10,000 times by randomly replacing words with a masking token (the unknown word token). For each perturbation, I first draw the number of words to be replaced $n \sim Uniform[1, N_i]$, where N_i is the number of words in the report. Then, I draw the n words to be replaced from the report text with equal probability. This gives each word a probability of $\frac{N_i+1}{2N_i} \approx 0.5$ to be replaced.

2. Use the neural network to predict the outcome for the perturbed documents.

3. Fit a weighted Lasso regression to the perturbed data and predicted outcomes. Use the regularization path to find the 20 most important features in the report (Efron et al., 2004). Examples are weighted by the cosine similarity between the perturbed example and the original text.

4. Store the regression coefficients (attribution scores) of the selected features.

Fig. 3.7 shows the estimated attribution scores for the most common words in a sample of 20,000 reports. A human reader would likely spot a difference in polarity between the word with positive and negative average attribution scores. Words with positive attribution score include words with positive polarity such as *strong*, *increased*, or *raised*, while *sluggish* is one of the most prominent words with negative attribution. To the extent that abnormal returns are at least partially realized on or around the publication day, this finding is in line with Huang et al. (2014), who document that the stock market reacts positively to analyst reports with positive sentiment and negatively to reports with negative sentiment. The attribution scores also reveal that certain topics are typically associated with negative or positive returns by the neural network. Among the top positive words are financial reporting-related variables such as *sales* and *revenue*, as well as *quarter*. Among the top negative words are terms that are typically used to express opinions such as *recommend*, *view*, and *believe*. All of the top terms appear to be firm-specific. Neither the positive nor the negative terms contain any terminology that is obviously related to macroeconomic conditions.

The plots reveal some limitations of the LIME procedure. Among the selected words are different forms of the word *are*. These words are unlikely to carry any return-relevant information. The LIME algorithm identifies words that, if removed from the given report, significantly change the prediction of the model. Therefore, the algorithm might select words that are important for the learned linguistic or syntactical interpretation of the surrounding words without being informative by themselves. If true positive terms are more reliant on the syntactical context than negative ones, or if the removal of a term creates false negatives, the term will appear to have a positive attribution and vice versa. It is reassuring that only a small fraction of the most common 50 positive and negative words appear to fall into this category.

The plots show the 50 most important positive (left panel) and negative (right panel) words identified by the LIME procedure as outlined in Section 3.7.1. Importance is measured as the probability of a word being selected by the lime procedure conditional on appearing in the report. Font size reflects the importance of each word. Font color reflects the average attribution, where green reflects positive attribution and red reflects negative attribution. Weaker colors reflect lower absolute attribution. Prepositions, determiners, and words that appear in less than 0.5% of the LIME explanations are excluded.

Figure 3.7: LIME word clouds

3.7.2 The role of industry-specific information

The LIME procedure allows us to study to which extent the HAN picks up universal predictors versus predictors that are only relevant for a subset of firms. To do so, I repeat the procedure outlined in Section 3.7.1 for different subsets of reports. I assign firms to five industries using SIC codes and the five Fama-French industry definitions from Ken French's website. For each industry subset, I train a separate LIME model using the same HAN that was learned based on all available data. Fig. 3.8 shows the most important words for each industry. A number of terms appear to be important in all five industries, in particular verbs and adjectives with strong polarity such as *raise[d]*, *solid*, and *strong*. Other terms appear to be industry-specific, for example, *campaign* in the Consumer industry or *subscribers* in the Business Equipment, Telephone, and TV Transmission sector. The Healthcare, Medical Equipment, and Drugs industry stands out for containing a large number of terminology related to partnerships, such as *partnership*, *partnering*, or *collaborations*. All of the partnership-related terms have a negative attribution, i.e. the HAN tends to predict a negative announcement return following the mentioning of partnerships.

The plots show the 50 most important words identified by the LIME procedure as outlined in Section 3.7.1 for different industries. Firms are assigned to industries using SIC codes and the five industry definitions of Fama and French. Each word cloud shows the most important words for each industry, where importance is measured as the probability of a word being selected by the lime procedure conditional on appearing in the report. Font size reflects the importance of each word. Font color reflects the average attribution, where green reflects positive attribution and red reflects negative attribution. Weaker colors reflect lower absolute attribution. Prepositions, determiners, and words that appear in less than 0.5% of the LIME explanations for a given industry are excluded.

Figure 3.8: LIME word clouds for different industries

3.7.3 Horse race

In this section, I compare the predictive power of the HAN model to several simpler prediction models: a fixed-effect model, a dictionary-based sentiment model, and an Elastic net model.
The fixed-effects model uses the prevailing historic mean announcement return for each firm as the prediction for the next announcement. In other words, the fixed-effects model estimates the mean announcement return for each firm in the training set and extrapolates this mean return to make out-of-sample predictions.
The dictionary-based sentiment model uses the Loughran and McDonald (2011) sentiment dictionary to assign a sentiment score to each report. Each word in the report is assigned a sentiment score of -1, 0, or 1 based on the sentiment dictionary. The report sentiment score is the average word sentiment score. I fit a linear regression model to predict announcement returns from the average report sentiment in the training sample and then use the fitted linear model to predict out-of-sample announcements returns.
The Elastic net also takes a linear bag-of-words approach. Words are represented as word counts, and the prediction is based on a linear regression of announcement returns on these word counts. The large vocabulary size relative to the number of examples and a high chance of collinear features makes a standard OLS approach infeasible. Therefore, I apply both an L_1 and a L_2 penalty on the coefficient matrix, a regularization method that is often referred to as Elastic net (Zou and Hastie, 2005). I select the regularization parameters through an exhaustive grid search. The exact specification of the benchmark models can be found in Appendix B.2.

Table 3.13 shows out-of-sample R^2 for the HAN model and the three benchmark models. Both the fixed effects model and the sentiment dictionary model perform worse than the historical mean benchmark. The fixed-effects model has an out-of-sample R^2 of nearly -6%, which suggests that announcement returns do not exhibit strong firm-fixed effects and that the HAN does not simply pick up persistent firm effects. The sentiment dictionary model has a marginally negative out-of-sample R^2 as well. Previous studies have shown that the stock market reacts to the sentiment in the reports upon publication (Huang et al., 2014), and that sentiment predicts analyst forecast errors (). The finding here suggests that the stock market fully incorporates the sentiment information leading up to the next announcement. The Elastic net has an out-of-sample R^2 of 0.11%. This suggests that even simple machine learning approaches are able to capture some of the return predictability in the analyst report dataset. It also provides a useful benchmark for the HAN model. The out-of-sample R^2 of the HAN model is four times higher than that of the Elastic net. The key difference between the HAN model and the Elastic net is that only the former takes into account contextual information. The large outperformance of the HAN model emphasizes the importance of this contextual information.

	HAN	Sentiment dictionary	Elastic net	Fixed effects
$R^2_{oos,mean}$ (%)	0.44	-0.00	0.11	-5.94
	(10.75)	(-3.41)	(8.13)	(-32.96)

The table shows out-of-sample R^2 for different prediction models for the 2004-2019 period based on biennial model updating. The HAN model is the neural network defined in Section 3.2. *Sentiment dictionary* makes return predictions based on a regression of returns on an analyst sentiment score using the Loughran and McDonald (2011) sentiment dictionary. *Elastic net* forms return predictions from analyst report texts using a bag-of-words elastic net regression. *Fixed effects* predict the prevailing historical average announcement day return for each firm. Diebold-Mariano test statistics are shown in parenthesis. The Diebold-Mariano test statistic for the difference between the HAN and its closest contender, the Elastic net, is 7.14 ($p < 10^{-12}$).

Table 3.13: Out-of-sample performance of different prediction models

3.8 Conclusion

I propose a new way to study stock return predictability that utilizes the vast amount of textual data that is available to investors. I show that a deep neural network that takes into account contextual information when interpreting individual words and optimally aggregates words and documents outperforms linear natural language processing techniques by a factor of four in terms of out-of-sample R^2.

I use the proposed methodology to study stock returns on earnings announcement dates. I find that market reactions to earnings announcements are predictable and that my findings are in line with biased cash flow expectations and not in line with common risk-based explanations. I show that the profitability factor loads strongly on an announcement return-prediction portfolio, suggesting that biased cash flow expectations play an important role in explaining this anomaly.

My paper highlights the power of big data and machine learning techniques to study classic questions in finance. Machine learning, in particular deep neural networks, allows researchers to tap unused data sources and to model complex interactions between input variables that might play an important role in explaining asset returns. While my paper focuses on individual stock return predictability on earnings announcement dates, a similar approach could be used to study long-horizon predictability or the predictability of aggregate returns, as well as higher moments or corporate decisions. The hierarchical attention network used in this paper can be readily applied to any source of textual data, for example, newspaper articles or corporate filings. The presented attention mechanism is particularly valuable in scenarios where multiple documents are linked to a single outcome variable and there does not exist an obvious aggregation method.

Chapter 4

Are Technological Innovators Priced Differently? Patent Intensity and Stock Returns, 1926-2021

4.1 Introduction

Technological innovation has long been proposed as a primary driver of economic growth (Schumpeter, 1911).[1] The public-market valuations of innovators drive costs of capital for both public and private firms,[2] influencing the viability of technology-driven growth. Leading models of the cross-section of stock returns such as Fama and French (1993, 2015), and Hou et al. (2015) do not explicitly account for technological innovation. The originators of these models invoke steady-state or static firm valuation models to obtain fundamentals-related pricing factors such as those based on the market/book ratio (Tobin's q), the capital investment rate, and profitability.[3] We ask what, if anything, do these steady-state-motivated models miss when applied to portfolios of technologically innovative firms?

We show that technological innovators and non-innovators are priced differently. We propose a new measure of technological innovation, patent intensity, given by the ratio of a firm's number of patents received to market capitalization. This simple new measure extends back to 1926, is not reliant on accounting data, and produces a significant spread in returns. Alpha remains after

[1] See also Solow (1957), Romer (1986, 1990), and Aghion and Howitt (1992).

[2] See Gompers et al. (2008).

[3] See Fama and French (1995) equation 2, Fama and French (2015) equation 3, Hou et al. (2015) equation 1. Berk (1995) provides a related valuation identity motivating size-related anomalies. The market/book ratio as a driver of investment is developed in Tobin (1958).

accounting for standard fundamentals-based factors. Innovators are not penalized for lack of profitability or high investment to the same degree as non-innovators. On average, innovative firms are larger than other firms, but the most innovation-intensive firms tend to be smaller firms with high growth and low profitability. Benchmarked to common asset pricing models, these firms have abnormally high returns.

We contribute to an important existing literature on technological innovation and the stock market that has for example examined the role of research and development expenses, patenting, and explicit models of technological progress. See, for example, Lev and Sougiannis (1996), Chan et al. (2001), Eberhart et al. (2004), Gu (2005), Cohen et al. (2013), Hirshleifer et al. (2013), Kogan et al. (2017), Hirshleifer et al. (2018), Bena and Garlappi (2020), and Kelly et al. (2021). We add to this literature by developing a new measure of innovation intensity based on patenting, and documenting the effectiveness of standard cross-sectional pricing models for technological innovators. An important property of innovation is its persistence. Portfolios formed on patent intensity have low turnover, and their return spread lasts ten full years following portfolio formation. This presents a significant challenge to asset pricing models.

The key to pricing portfolios sorted on patent-intensity is the expected growth factor of Hou et al. (2021, HMXZ). We show that in models with investment and profitability factors but not expected growth, abnormal returns of portfolios of innovative firms are statistically and economically significant for a full decade following formation. Including the expected growth factor eliminates abnormal returns of patent-intensity sorted portfolios at nearly all horizons.

Risk dynamics of technological innovators in the decade following formation tell a compelling economic story. Innovators load heavily on expected growth immediately following formation, and over time their expected growth loadings fall. Even a decade after formation, the growth loadings of innovators significantly exceed those of non-innovators. Investment loadings of innovators are initially somewhat aggressive, but become even more so for two to three years following portfolio formation. Investment loadings remain higher than non-innovators for a full decade. Finally, innovators show extremely weak profitability loadings immediately after formation, but strengthen substantially over the following decade.

Previous work has shown positive abnormal returns for R&D sorted portfolios (Chan et al., 2001, HMXZ). We differ from these studies in several ways. First, these studies focus on firms with strictly positive R&D data, which excludes firms with zero or missing R&D, comprising on average half of firms by market capitalization. We use patent intensity, which allows us to unambiguously categorize non-innovators. Further, patent intensity can be measured over much longer samples, since it does not rely on accounting data. We show however that patent intensity and R&D intensity are closely related over the period over which they can both be measured. Our study focuses exclusively on the pricing of innovators, and we study the drivers and dynamics of risk and return for innovative firms.

Innovating firms have played an important role in the US stock market for at least one century. While the firms and industries that were dominating the innovative landscape have varied substantially over time, from manufacturing firms in the mid-20th century to computer and information technology companies in the most recent two decades, the overall share of innovators in the US stock market has remained remarkably constant. Throughout the 1926-2021 sample period, innovators have accounted for approximately 40-80% of the total US market capitalization. Therefore, the pricing of these firms is not only highly relevant for our understanding of asset pricing models, but is also critical for capital allocation and ultimately economic growth.

4.2 Technological innovators and patent intensity

This section describes the patenting activity of publicly listed firms in the United States from 1926. We define our main variable, patent intensity (PI), and show the characteristics of more and less patent-intensive firms. Patent intensity is highly persistent.

4.2.1 Patent data and innovative firms

The United States Patent & Trademark Office (USPTO) is the source of complete data for all patents granted. The USPTO provides downloadable text data starting in 1976.[4] For the universe of all patents filed between 1926-1975, Kelly et al. (2021) provide cleaned and tabulated patent data created from USPTO image files.[5] In combination, these two sources provide full coverage of all U.S. patents issued from 1926-2021. We link patents to publicly listed companies using CRSP permno-patent links from Kogan et al. (2017).[6]

U.S. firms with common stock traded on NYSE, AMEX, or Nasdaq are important contributors to overall U.S. patenting. Each calendar year, starting in 1926, we calculate the share of patents for all CRSP assignees (includes foreign firms), as well as the share of patents for all U.S. listed firms with common stock (CRSP shrcd is 10 or 11). Fig. 4.1, Panel A shows the logarithm of the patent counts for each group (universe, all CRSP, and listed U.S. common stock). Panel B shows the shares of all CRSP assignees and U.S. listed common stock. The wedge between all CRSP assignees and U.S.-listed common stock assignees toward the end of the sample is due to the growing importance of cross-listed foreign firms that receive U.S. patents. Patenters with U.S.-listed common stock are important throughout the sample, with a share of overall patenting ranging from twenty to forty

[4] https://www.uspto.gov/patents/search.
[5] https://github.com/KPSS2017/Measuring-Technological-Innovation-Over-the-Long-Run-Replication-Kit.
[6] https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Replication-Kit.

percent throughout most of the sample. These firms are key building blocks of empirical asset pricing studies. Publicly listed patenters are also economically important because they provide a broad investor base access to equity in technological innovators, and because daily updated prices reflect a market view of the value of innovation.

From the standard CRSP sample of all common stock (shrcd is 10 or 11) traded on NYSE, AMEX, or Nasdaq, each year on June 30 we classify firms as "innovators" or "non-innovators" based on whether they received a patent in the prior 12-month period. The USPTO publishes its Official Gazette every Tuesday with information on patents granted that day, so patent information is immediately available to market participants.[7]

Relative to other measures of innovation such as accounting-based measures of R&D, a patent-based classification of innovators is appealing because it is based on a standardized and tangible legal claim. Patent data is not subject to the reporting practices of individual firms, and reporting cannot be missing or delayed. Patents measure the output of the innovation process, whereas R&D measures the input. Our choice of a twelve-month lookback period for measuring innovation is simple and convenient. Our results are robust to variations such as measuring patenting activity over the prior calendar year, or to using longer lookback periods, such as patents received over a three-year window. Choosing a one-year window for our main results ensures that our results about the persistence of patenting activity are not artificially driven by overlapping measurement windows. To avoid any inconsistencies with assigning patents to newly listed firms, we drop firms from our analysis that have less than a twelve-month history in the CRSP data.[8]

Fig. 4.1, Panels C-F show that despite some sharp fluctuations in the percentage of innovators versus non-innovators over time by firm count, the percentage of innovators by market capitalization is much slower-moving, and appears mean-reverting. The share of innovators by number of firms (Panel D) ranges from about twenty to fifty percent throughout the sample. By market capitalization, the share of innovators generally ranges from fifty to seventy-five percent, consistent with innovators being larger. All of our main results use value-weighted portfolios, so the more stable market-capitalization weighted shares of innovators versus non-innovators are most relevant.

A coarse example shows that the sector composition of innovative and non-innovative firms varies considerably over time. Each year we assign all CRSP firms to one of ten Fama-French industries, each of which can be thought of as a sector. Fig. 4.2, Panel A shows sector allocations over time from the market-capitalization weighted portfolio of all innovative firms. Panel B shows the sector allocations for the market-capitalization weighted portfolio of all non-innovative firms.

[7] https://www.uspto.gov/learning-and-resources/official-gazette.

[8] Links from patent assignees to CRSP firms are reliable, but linking assignees to firms before they become public is more challenging.

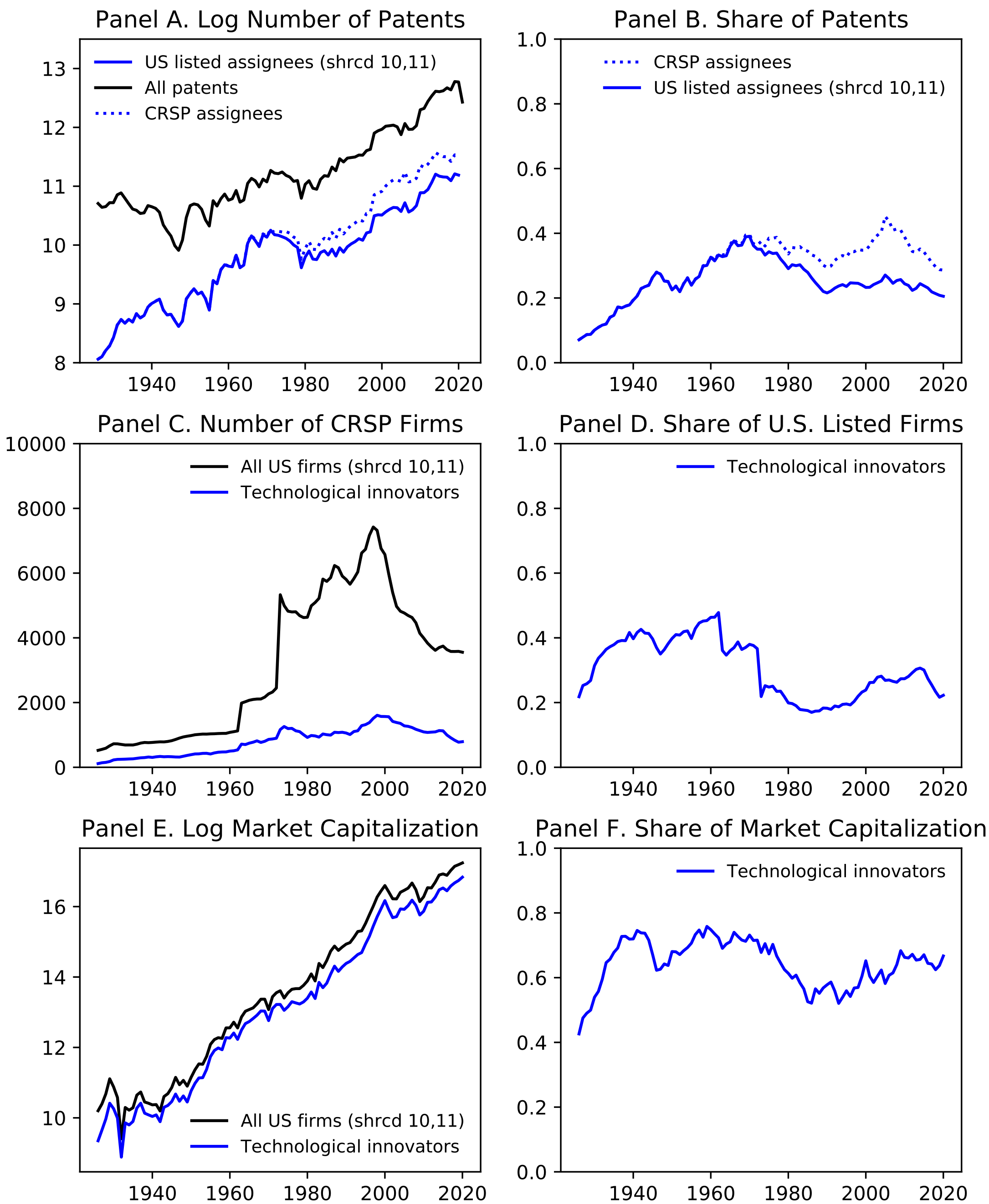


Figure 4.1: Patents, US-listed firms, technological innovators and market capitalization

The sector allocations change considerably throughout the sample. For example, the importance of manufacturing and consumer durables in the innovator portfolio decreases over time, while the importance of business equipment and healthcare increases. Technological innovation concentrates in different sectors of the economy throughout our sample.

4.2.2 Patent intensity

Starting in 1926, on June 30 of every year we calculate for every firm in the CRSP sample the ratio of patents received in the prior 12-month period divided by current CRSP market capitalization. This is our measure of patent intensity (PI). All results in the remainder of the paper are robust to reasonable alternative choices such as measuring patent intensity at the end of the prior calendar year, or over three-year windows. We choose a one-year window because of its simplicity and because a one-year window does not generate mechanical persistence in the measure. Scaling by market capitalization is a natural choice and makes PI comparable to prior measures such as the book-to-market ratio, which can be thought of as a measure of asset intensity, or R&D to market capitalization. Conceptually, purchasing firms with high PI allows an investor to obtain the most concentrated exposure to recent patenting activity with the least dollar investment.

To begin our characterization of patent intensity, each year we sort firms into three groups. Group zero has no patents in the prior twelve-month period, and are the "non-innovators" described previously. We note that, unlike other variables, there is no issue of "missing data" with patents. For example, missing data is common in R&D data, and many researchers (e.g., Chan et al., 2001) discard from analysis firms with missing R&D data. Ambiguity caused by missing data is not an issue with patents. We distinguish between low- and high-intensity patenters, each year dividing all innovators at the median positive PI breakpoint, forming two equal-sized groups by firm count.

Table 4.1 provides descriptive statistics for the three groups, showing important differences. Panel A shows the average contributions of each of the three groups to firm counts, total market capitalization, and past and future patenting. Most firms (68% on average) are non-patenters. Nonetheless, the 32% of patenting firms contribute the majority of market capitalization, 65% in an average year. The concentration of market capitalization is even more noticeable if we look at the high- and low-PI groups. The low PI group, while only 16% of firm count, contributes 54% of market capitalization. The high-PI group is again 16% by firm count, but only 11% by market capitalization.

It would be a tremendous mistake to conclude that the high-PI group is inconsequential because of its small market capitalization. This group owns on average 62.5% of the universe of patents created by public firms in the prior year. The majority of innovation has occurred within this group. Moreover, this is not just *ex post* selection. The high-PI group also contributes 60% of the patents

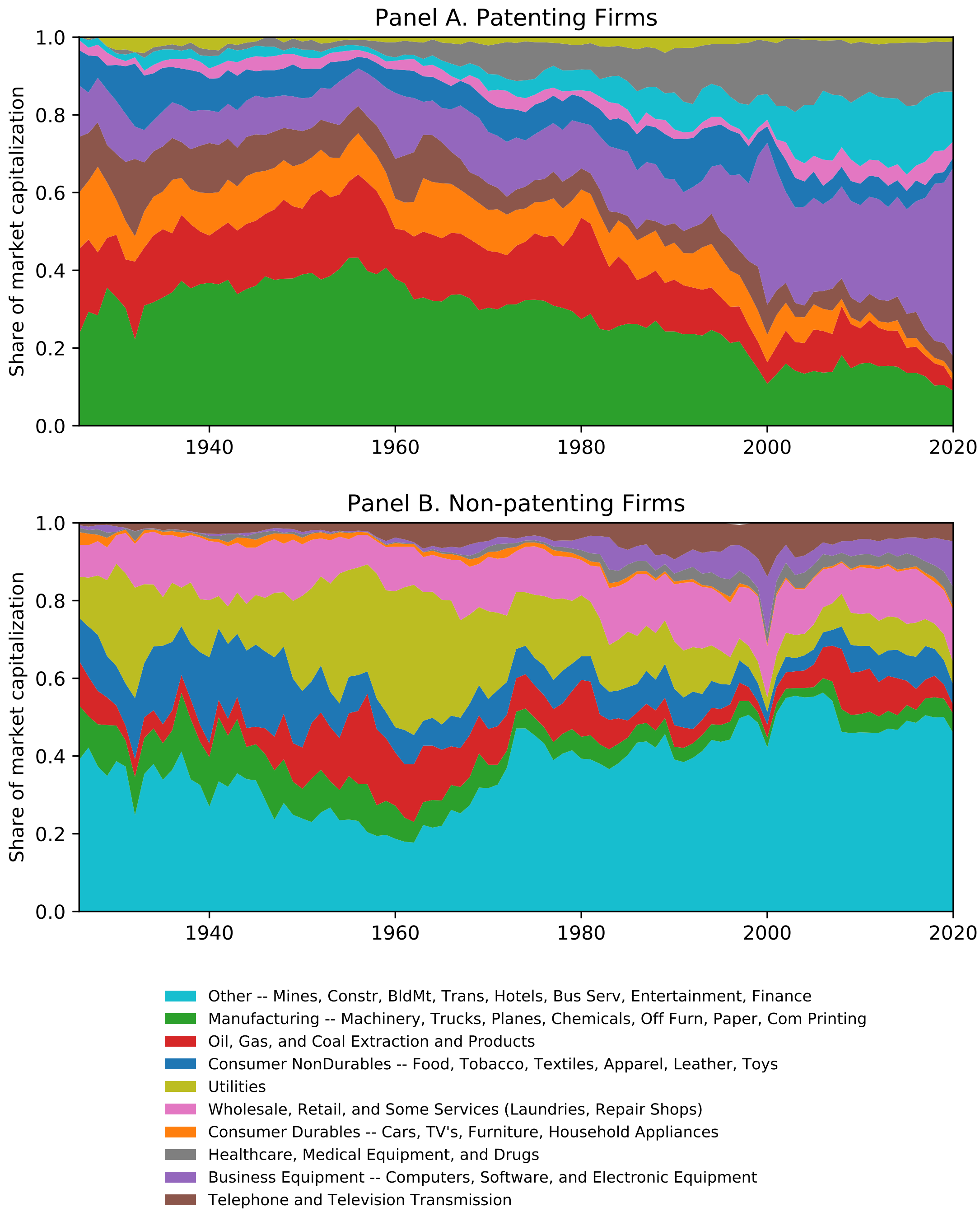


Figure 4.2: Sector composition of patenting and non-patenting firms by market capitalization and Fama-French 10 industries

	Non-patenting	Low PI	High PI
Panel A. Portfolio shares			
Share of firms	0.682	0.159	0.159
Share of cap	0.349	0.538	0.113
Share of patents	0.000	0.375	0.625
Share of patents (next year)	0.012	0.391	0.597
Share of patents (next 3 years)	0.014	0.405	0.580
Share of patents (next 5 years)	0.017	0.418	0.565
Panel B. Portfolio variables			
CRSP age mean	13.313	20.283	15.422
CRSP age median	11.711	17.685	12.998
BM mean	1.573	0.801	1.130
BM median	0.998	0.662	0.909
Investment mean since 1963	0.137	0.164	0.090
Investment median since 1963	0.075	0.091	0.043
Profitability mean since 1963	0.164	0.257	0.095
Profitability median since 1963	0.210	0.264	0.165

This table shows descriptive statistics and characteristics of firms sorted on patent intensity PI, patents received in the prior year divided by market capitalization. Firms are sorted every year at the end of June into three groups. The first group consists of non-patenting firms ($PI = 0$). Remaining firms are split equally into two groups, low and high PI. In panel A, share of firms is the portfolio's percentage of all companies, share of cap is the portfolio's share of total market capitalization and share of patents is the portfolio's share of all patents at the time of sorting or as indicated. Panel B shows descriptive statistics based on information available at the time of sorting. Mean and median indicate whether the value is from cross-sectional mean or median, respectively, before averaging across years. Age is calculated from the stock's first appearance in CRSP. Investment and profitability are available only since 1963. For all numbers, we first calculate the annual percentages (or mean and median as indicated) and then average across years from 1926 to 2021, or as indicated.

Table 4.1: Patent intensity (PI) and firm characteristics

granted to public firms in the next year, 58% of patents granted in the next three years, and 56.5% of patents in the next five years. Patenting activity is very persistent, and with a relatively small allocation of equity capital (11.3% of total market capitalization), one can purchase the majority of not only recent but also future five-year ahead public market patenting activity.

Table 4.1, Panel B shows further characteristics of the three groups. Non-innovators are younger on average and by median than innovators. This may seem surprising given the stereotype of young firms as innovators, but average age also relates to death rate, which we explore further below. Among innovators, high-PI are younger than low-PI, consistent with intuition. The B/M ratio is a traditional measure of "value", and non-innovators have the highest B/M ratios. Interestingly, high-intensity innovators appear to be more value-like than low-intensity innovators. This should not be too surprising, since both PI and B/M have market capitalization in the denominator. We can usefully think of PI as a measure of technological-innovation value or the most cost-efficient way to purchase patenting activity. Considering investment and profitability, low-intensity innovators have the highest investment rates in traditional assets and the highest profitability. High-intensity innovators have both the lowest investment in traditional assets and the lowest profitability.

Table 4.1 thus shows that technological innovation intensity captures important differences across firms. The archetype of a non-innovator is a modestly sized, perhaps shorter-lived value firm with moderate investment and profitability. Low patent-intensity firms are larger, longer-lived firms that appear successful, investing in traditional assets and maintaining high profitability, and appearing as "growth" by low B/M. High-intensity innovators are young and small, somewhat counterintuitively appear as "value" by the B/M measure, invest the least in traditional assets and have the lowest profitability, but produce the lion's share of technological innovation among listed public firms. Because of the important differences across these categories of firms, we anticipate a meaningful challenge for traditional pricing factors such as size, value, investment, and profitability to price PI-sorted portfolios.

The NASDAQ exchange has a reputation as a listing place for technological innovators,[9] and Table 4.2 shows the importance of NASDAQ for patenting. Panel A shows that by firm count, most of the firms on NASDAQ are non-innovators. But by market capitalization, NASDAQ has been shifting more and more to be represented by low-intensity innovators. Panel B shows the contribution of NASDAQ firms to the PI-sorted portfolios. In recent years, more than fifty percent of the cap weight and forty percent of the patents of low-intensity innovators have come from NASDAQ. For high-intensity innovators, more than forty percent of both the cap weight and patents come recently from NASDAQ.

Finally, Table 4.3 shows in Panel A average transition probabilities across the three PI-sorted

[9]See for example Schwert (2002); Pástor and Veronesi (2006, 2009).

	Non-patenting	Low PI	High PI
Panel A. NASDAQ Composition (columns add to one)			
Share of NASDAQ firms since 1973	0.797	0.074	0.129
Share of NASDAQ firms since 2000	0.720	0.108	0.171
Share of NASDAQ firms since 2015	0.732	0.105	0.164
Share of NASDAQ cap since 1973	0.552	0.361	0.087
Share of NASDAQ cap since 2000	0.309	0.548	0.143
Share of NASDAQ cap since 2015	0.252	0.621	0.127
Share of NASDAQ patents since 1973		0.303	0.697
Share of NASDAQ patents since 2000		0.343	0.657
Share of NASDAQ patents since 2015		0.412	0.588
Panel B. NASDAQ Shares of Column (1-entry is non-NASDAQ)			
Firms from NASDAQ since 1973	0.606	0.352	0.604
Firms from NASDAQ since 2000	0.598	0.471	0.742
Firms from NASDAQ since 2015	0.587	0.472	0.732
Cap from NASDAQ since 1973	0.219	0.155	0.205
Cap from NASDAQ since 2000	0.229	0.277	0.371
Cap from NASDAQ since 2015	0.262	0.407	0.441
Patents from NASDAQ since 1973		0.185	0.218
Patents from NASDAQ since 2000		0.356	0.360
Patents from NASDAQ since 2015		0.517	0.407

This table reports the portfolio shares and composition of NASDAQ-listed companies across portfolios of non-patenting, low-PI, and high-PI firms as defined in notes of Table 4.1. The share of NASDAQ firms in panel A is the portfolio's average percentage of all NASDAQ-listed companies in the specified time period. The share of NASDAQ cap and the share of NASDAQ patents are equivalently portfolio's average percentages for market capitalization and patents, respectively, of NASDAQ listed companies. Firms from NASDAQ in panel B shows the average percentage of the firms in the portfolio over the indicated time period that are listed on NASDAQ. Cap and patents from NASDAQ are defined equivalently for market capitalization and patents, respectively, of NASDAQ-listed companies. For all numbers, we first calculate the annual percentages and then average across the indicated time period.

Table 4.2: Technological innovators on NASDAQ

portfolios as well as exit, at horizons of one, three, and five years. For comparison, Panel B shows similar transition probabilities for the traditional measure of growth, the M/B ratio. To enhance comparison, we set the breakpoints for the M/B sort in Panel B identically on a year-by-year basis to the breakpoints for the PI sorts in Panel A.[10]

[10]Transition probabilities in Panel B are calculated conditional on not having a negative or missing book value. Missing or negative book values are not trivial, 12% of the sample on average, which is a general difficulty for accounting-based characteristics that does not apply to patent intensity.

Panel A. PI-sorted portfolios					Panel B. M/B-sorted portfolios					
	Non-patenting	Low PI	High PI	Out		Low M/B	Medium M/B	High M/B	Missing	Out
Transition probabilities over 1 years										
Non-patenting	86.8	4.6	2.6	6.1	Low M/B	85.6	6.5	1.6	3.6	2.7
Low PI	17.7	67.0	12.7	2.6	Medium M/B	35.0	43.8	16.0	3.3	2.0
High PI	12.8	12.0	71.1	4.1	High M/B	9.8	19.9	63.0	5.2	2.1
					Missing	9.4	2.3	5.2	57.3	25.9
Transition probabilities over 3 years										
Non-patenting	75.9	5.2	2.8	16.1	Low M/B	73.6	7.7	3.0	3.8	11.9
Low PI	17.4	58.7	15.9	8.0	Medium M/B	43.6	28.1	15.2	3.7	9.4
High PI	13.2	15.0	60.0	11.8	High M/B	20.9	20.4	43.8	5.4	9.6
					Missing	14.9	4.2	5.6	39.0	36.2
Transition probabilities over 5 years										
Non-patenting	67.5	5.5	2.9	24.1	Low M/B	65.6	7.8	3.5	3.6	19.6
Low PI	17.1	53.6	16.7	12.5	Medium M/B	45.0	22.2	13.5	3.5	15.8
High PI	12.8	16.4	52.4	18.4	High M/B	25.5	18.5	35.1	4.8	16.1
					Missing	17.6	5.1	5.2	28.7	43.4

Panel A shows the transition probabilities between portfolios of stocks sorted by *PI* as described in notes of Table 4.1 over 1, 3, and 5 years. Rows specify the initial portfolio and columns the portfolio of the stock after the indicated time period. Column "out" reports the probability of a stock disappearing from the data. The probabilities in each row are conditional, indicate the probability of moving from the initial portfolio (rows) to the destination portfolio in columns (or out), and sum up to 1 across the columns. Panel B shows the equivalent for market-to-book (M/B)-sorted portfolios and defines an additional "Missing" portfolio consisting of firms with negative or missing market-to-book ratio. To allow a fair comparison of the transition probabilities of the PI-sorted portfolios with the transition probabilities of M/B-sorted portfolio, the M/B-sorted portfolios are based on the same percentiles as PI-sorted portfolios: each year, we calculate the percentages of firms in each of the three PI-sorted portfolios and use these percentages to categorize stocks by M/B. The unconditional probabilities (shares) of non-patenting, low-PI, and high-PI portfolios are 68.2%, 15.9% and 15.9%, respectively (see Table 4.1). These probabilities apply also to the M/B-sorted portfolios for stocks with non-missing M/B. 88% of stocks have non-missing M/B, the remaining 12% have missing M/B. Accordingly, the unconditional probabilities of the four M/B-sorted portfolios are: 68%*88%=59.4% (low M/B), 15.9%*88%=14% (medium M/B), and 15.9%*88%=14% (high M/B). Transition probabilities are calculated annually over period from 1926 to 2020. The presented transition probabilities are time-series averages.

Table 4.3: Transition probabilities of PI- vs. M/B-sorted portfolios

One key message from Table 4.3 is the high exit rate of non-innovative firms. Compared to the low-M/B firms in Panel B at one-, three-, and five-year horizons, the non-innovator versus low-M/B delisting rates are respectively 6.1 vs. 2.7%, 16.1 vs. 11.9%, and 24.1 vs. 19.6%. The high delisting rate of non-innovators helps to explain their low average age shown previously. Further, the majority of delistings are negative events, which are known to impact portfolio performance (Shumway, 1997).

A second key finding from Table 4.3 is the persistence of PI sorts. For every horizon, high-PI firms are considerably more likely to remain high-PI firms in the future than are high-M/B firms. Low-PI is similarly more persistent than medium M/B. Comparing non-patenters to the low M/B firms, persistence is modestly higher at all horizons, but given the higher exit rates of non-patenters this persistence is still noticeable. Because of the persistence of the PI characteristic, we expect portfolio sorts to be relatively low-turnover.

4.3 Patent intensity and stock returns

In this section, we compare the stock returns of innovators and non-innovators. Innovators have higher returns than non-innovators, both in raw returns and after controlling for common risk factors. We show similarities in sorts on R&D intensity and patenting intensity, and demonstrate that controlling for expected growth is crucial to capturing the returns of innovating firms.

4.3.1 Stock returns of patent intensity portfolios

We use two samples in this subsection. The first, full sample, begins in July, 1926. The second sample begins in July, 1963 to accommodate performance analysis with the Fama-French five-factor model, whose investment and profitability factors begin that month.

In the full sample, the portfolios are exactly as in the prior section: non-innovators (no patents, denoted portfolio “0”), low-intensity innovators (lower half of PI sort, portfolio “1”), and high-intensity innovators (upper half of PI sort, portfolio “2”). The 1963-2021 period eliminates early years with much smaller numbers of firms, so we sort innovators into bins of four bins with equal numbers of firms. We label these 1-4. The sorts thus appear numbered as tercile or quintile sorts, but portfolio zero always corresponds to non-innovators ($PI = 0$), and positive-numbered portfolios are innovators ($PI > 0$) sorted by PI into bins with equal numbers of firms. Portfolio HL is a zero-cost portfolio with a short position in the non-patenting portfolio “0” and a long position in the highest PI portfolio. Table 4.4 shows value-weighted monthly excess returns (Panel A), CAPM regressions (Panel B), Fama-French three-factor regressions (Panel C), and Fama-French five-factor regressions (Panel D). The left-hand side of the table shows full-sample results and the right-hand side shows the 1963-2021 sample.

In Table 4.4, Panel A, the annualized average excess returns (monthly returns multiplied by twelve) increase monotonically across portfolios in the full sample from 7.68% for the non-patenting portfolio 0 to 11.79% for the high-PI stocks. The sample starting in 1963 confirms the increasing average excess returns across the more granular sort into five portfolios. The pattern is again monotonic with the exception of portfolio 1 having a slightly lower return than non-patenting portfolio. The HL portfolio earns economically and statistically significant returns of 4.1% over the full sample and 6.97% over the post-1963 sample.

The CAPM regressions in Panel B show that market betas are slightly increasing across the PI-sorted portfolios, but not sufficiently so to explain the excess returns of the high-PI portfolio. The HL alpha is 2.28% p.a. in the full-sample and 5.12% post-1963, both statistically significant. The FF3 regressions in Panel C show similar alphas – controlling for size and book-to-market factors does not substantially change our inference about portfolio performance. We do see that non-innovative firm loadings are consistent with small size and value. Among innovators, higher PI is associated with greater size loadings and somewhat more value than growth.

Despite the common description of the HML factor as value versus growth, the FF3 results could be consistent with HML playing dual contradictory roles for PI-sorted portfolios. Firms in the high-PI portfolio do a lot of patenting, which we naturally think of as a predictor of growth. At the same time, investors can acquire these firms with minimum employment of equity capital, which seems to indicate value. The value-growth paradigm faces the difficulty that value and growth do not seem to be opposites in a single dimension, but two distinct concepts. Value-growth can be effective in a low-dimensional factor model because it relates to and summarizes several other useful sources of variation, but in higher-dimensional models, HML becomes less informative as those other sources of variation are parsed explicitly (Fama and French, 2015). This difficulty can be seen in the PI sorts. Despite the very large variation in the types of firms in the PI-sorted portfolios, we see surprisingly little variation in the HML loadings. The HML factor cannot help to explain the returns of technological innovators.

The FF5 regressions in Panel D, which add investment and profitability factors, cannot resolve the mispricing of technological innovators. In fact, if anything the difficulties deepen. The profitability loadings align very strongly with the technological innovation sort, but in the opposite direction needed to explain the pattern of returns. High-PI firms have very negative profitability loadings, and non-innovators have positive profitability loadings. Higher profitability is supposed to earn a premium according to the profitability factor, but that pattern is reversed in the PI sorts. Investment loadings are not statistically significant but align in the right direction to help explain returns. The net effect is that the five-factor model produces a stronger alpha-sort than the CAPM

	1926-2021				1963-2021					
	0	1	2	HL	0	1	2	3	4	HL
Panel A. Excess returns										
Excess return	7.68***	8.41***	11.79***	4.1***	6.58***	6.18***	8.62***	9.56***	13.54***	6.97***
	(3.82)	(4.41)	(4.75)	(3.84)	(2.95)	(3.14)	(3.92)	(3.8)	(4.1)	(3.43)
Panel B. CAPM										
Constant	-0.44	0.5	1.83**	2.28**	-0.34	-0.24	1.59**	1.62	4.78**	5.12***
	(-0.93)	(1.62)	(2.29)	(2.32)	(-0.51)	(-0.46)	(2.32)	(1.55)	(2.52)	(2.61)
Mkt-RF	0.98***	0.95***	1.2***	0.22***	0.99***	0.92***	1.01***	1.14***	1.25***	0.26***
	(67.76)	(92.35)	(42.93)	(8.23)	(55.95)	(70.49)	(59.31)	(42.67)	(26.31)	(5.37)
R^2	0.95	0.97	0.89	0.16	0.93	0.94	0.9	0.82	0.67	0.07
Panel C. Fama-French 1993										
Constant	-0.89**	0.78***	1.45*	2.35**	-1.36***	0.41	1.83***	1.23	3.63**	4.99***
	(-2.23)	(2.96)	(1.84)	(2.35)	(-2.83)	(0.97)	(2.64)	(1.17)	(2.14)	(2.6)
Mkt-RF	0.94***	0.99***	1.15***	0.2***	1.0***	0.94***	1.0***	1.08***	1.12***	0.12*
	(86.04)	(122.77)	(42.37)	(6.24)	(68.43)	(96.36)	(52.54)	(31.78)	(20.54)	(1.9)
SMB	0.08**	-0.12***	0.25***	0.17*	0.09**	-0.2***	-0.01	0.3***	0.7***	0.61***
	(2.41)	(-9.47)	(3.62)	(1.79)	(2.06)	(-15.68)	(-0.26)	(3.73)	(5.78)	(3.84)
HML	0.14***	-0.06***	0.06	-0.09	0.22***	-0.1***	-0.06*	0.01	0.09	-0.13*
	(5.8)	(-4.38)	(1.2)	(-1.43)	(7.21)	(-5.48)	(-1.75)	(0.31)	(1.4)	(-1.67)
R^2	0.96	0.98	0.91	0.2	0.95	0.96	0.9	0.85	0.76	0.25
Panel D. Fama-French 2015										
Constant					-1.76***	0.13	2.11***	2.02**	4.94***	6.71***
					(-3.75)	(0.3)	(3.03)	(1.97)	(2.96)	(3.54)
Mkt-RF					1.0***	0.95***	1.01***	1.08***	1.12***	0.12**
					(76.92)	(93.56)	(64.84)	(33.85)	(23.2)	(2.29)
SMB					0.13***	-0.18***	-0.05*	0.22***	0.57***	0.44***
					(5.41)	(-14.08)	(-1.73)	(3.87)	(7.38)	(4.71)
HML					0.21***	-0.08***	-0.08**	-0.08	-0.1	-0.31***
					(6.57)	(-3.69)	(-1.99)	(-1.18)	(-1.08)	(-2.81)
CMA					-0.03	0.02	0.08	0.14	0.24	0.27
					(-0.81)	(0.58)	(1.37)	(1.55)	(1.58)	(1.6)
RMW					0.13***	0.05***	-0.12***	-0.28***	-0.45***	-0.58***
					(3.23)	(2.63)	(-3.38)	(-3.83)	(-3.11)	(-3.47)
R^2					0.96	0.96	0.9	0.86	0.77	0.32

The table shows the average excess returns of PI-sorted portfolios in panel A and the results of regressing the portfolio returns on a constant and market excess returns, Fama-French 3 factors and Fama-French 5 factors in panels B, C, and D, respectively. Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios each year at the end of June. All portfolios are value-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, and the estimates of the average excess returns and constants are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is indicated in headings, i.e., 1926-2021 and 1963-2021. Data for Fama-French 5 factors is available from 1963. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.4: Patent Intensity sorts and performance, Fama-French factors

or three-factor models, with a highly statistically significant HL alpha of 6.7%.

One other item of note from Panel D is the abnormal negative five-factor performance of non-innovators (portfolio 0). This portfolio can be formed with a simple indicator variable, whether a firm received a patent in the last year or not. Though the magnitude of the alpha is economically modest, -1.76% per year, it is highly statistically significant. Non-innovators earn negative abnormal returns according to very standard benchmark models (see also Panel C in both samples).

The results presented in this subsection are robust to including a momentum factor as in the Fama and French (2018) six-factor model, as shown in Table C.3 in the appendix. Momentum loadings on the PI sorted portfolios are generally small and do not change alphas substantially.

4.3.2 Comparison with R&D intensity

Research and development expenditures and patents both capture aspects of the innovation process. R&D expenditures are an input to technological innovation, whereas patents are an output. While the success of research and development is uncertain, prior literature (e.g., Bound et al., 1982) shows that R&D expenses predict patenting. We therefore expect portfolios sorted on R&D to relate to portfolios sorted on patent intensity.

Following prior literature, we measure R&D intensity (RDI) on June 30 as the ratio of R&D expense (prior fiscal year) to CRSP market capitalization (calendar end of prior year) starting in 1975. Chan et al. (2001) first show a positive relationship between R&D expenses and returns. They scale R&D expenses by market capitalization and begin their sample in 1975. Although R&D data is available prior to 1975, in 1974 the FASB issued SFAS No. 2, which standardized and required accounting for R&D costs.[11] Hou et al. (2021) confirm a positive relationship between R&D and abnormal returns with standard factors in a sample extended to 2016.

Our measure of R&D intensity (RDI) is identical to the R&D to market equity variable used in prior literature, but we make one important change to methodology in the treatment of missing or zero R&D expenses. Both Chan et al. (2001) and Hou et al. (2021) include only stocks with positive R&D expenses, sorting into quintiles and deciles, respectively. Stocks with missing or zero R&D are excluded.[12] We treat stocks with missing R&D data in Compustat as having no R&D. Following the sorting methodology we use for patents, our portfolio zero comprises all stocks having zero or missing R&D expenses ("non-innovators"), and from the remaining firms with positive R&D expenses ("innovators") we sort into four bins by RDI with equal numbers of firms.

[11] See *Statement of Financial Accounting Standard No. 2: Accounting for Research and Development Costs* at https://fasb.org/referencelibrary. The impact of this change has been studied in the accounting literature. See Elliott et al. (1984).

[12] See Chan et al. (2001) notes to Table VI, page 2449, and Hou et al. (2020b) Appendix A.5.4, page 2104. See also Cohen et al. (2013).

Our approach to missing or zero R&D data is different but informative. First, as Peters and Taylor (2017) explain, SFAS No. 2 gives us reasonable confidence that firms with missing R&D expenses in Compustat after 1974 typically did not incur such expenses, i.e. can be treated as zero. Second, the identical treatment of our R&D sort with our patent sort gives greater comparability of results. Third, the effects of our treatment of R&D expenses can be checked *ex post*. If our portfolio zero of non-innovators with R&D looks similar to our portfolio of non-innovators with patents, where there is no missing data, then this gives confidence that treating the absence of R&D expenses as no R&D expenses is reasonable. Finally, including firms with zero or missing R&D expenses greatly expands the scope of our analysis. In the post-1975 period, firms with zero or missing R&D comprised 60-70% of the total universe by firm count, and 40-50% of the total universe by market capitalization, as shown in Fig. 4.3. Including these firms in our analysis therefore gives a useful check of the relationship documented in earlier literature on a broader sample.

Table 4.5 shows results for return performance of the RDI portfolios. Panel A confirms that firms with high RDI have higher returns than firms with low RDI. The average annual excess returns of firms in the highest RDI quartile are approximately 6.41% higher than those in the lowest quartile. The average excess return of firms with no research and development expenses, shown in portfolio "0", is slightly higher than that of firms in portfolio "1", but still substantially lower than the return of high RDI firms. Panel B shows risk-adjusted returns controlling for the Fama-French five factors. Here, the importance of separating low-R&D firms from no-R&D firms becomes evident. While low-R&D firms are correctly priced by the Fama-French five-factor model, no-R&D firms have a statistically significant negative alpha of -1.06% per year. This finding mirrors the results for the low-patenting versus no-patenting portfolios shown in the previous section.

Table 4.6 compares the high-minus-low RDI portfolio with the high-minus-low PI portfolio. Columns 1 and 3 show FF5 regressions for PI and RDI, respectively. Both HL portfolios load similarly on the Fama-French five factors, for example loading very negatively on profitability (-0.64 and -0.71 for PI and RDI, respectively), somewhat negatively on value (-0.36 and -0.24), and positively on investment (0.41 and 0.32 implying conservative investment in traditional assets). These results confirm that the two portfolios have similar risk exposures and returns. In columns 2 and 4, we test whether the RDI portfolio spans the PI portfolio and vice versa. Column 2 shows that the PI portfolio loads strongly on RDI (loading of 0.61), and the regression R^2 increases from 0.34 in column 1 to 0.56 in column 2. Alpha falls by approximately one-half from column 1 to column 2, leaving a significant abnormal return of 2.92% unexplained. Column 4 similarly shows that the PI portfolio explains significant variation in the RDI portfolio, and about two-thirds of the RDI alpha is eliminated with the remainder being statistically indistinguishable from zero.

The overriding takeaway from this analysis is that patenting intensity and R&D intensity, both

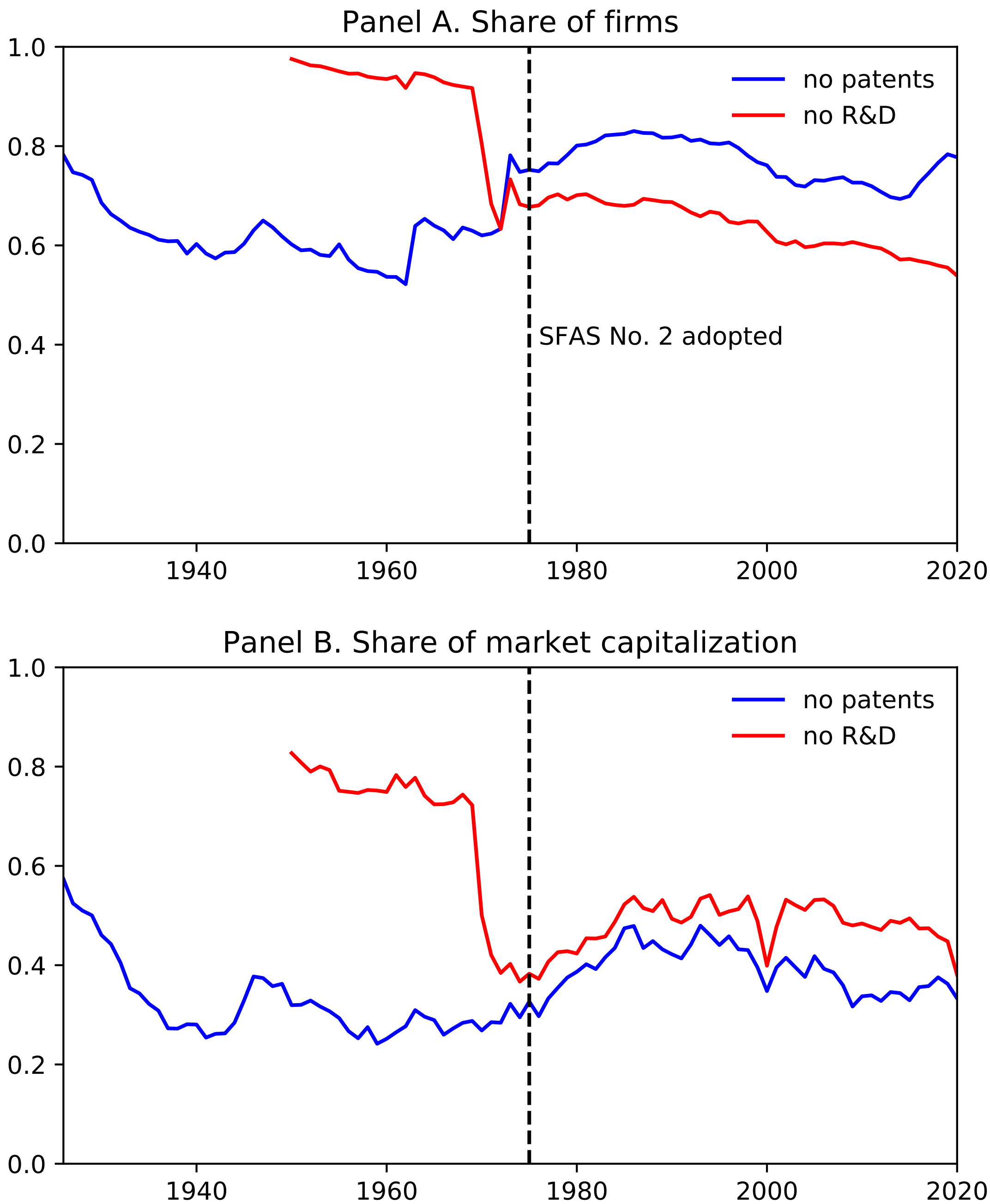


Figure 4.3: Fraction of non-R&D and non-patenting firms in the sample

	0	1	2	3	4	HL
Panel A. Excess returns						
Average	8.459***	7.127***	9.890***	11.227***	13.538***	5.078**
	(3.55)	(3.08)	(3.92)	(4.18)	(3.64)	(2.23)
Panel B. Fama-French 2015						
Average	-1.056**	0.209	1.973**	3.082***	3.681**	4.737**
	(-2.46)	(0.25)	(2.17)	(2.79)	(2.07)	(2.46)
Mkt-RF	1.005***	0.951***	1.019***	1.022***	1.133***	0.129***
	(92.55)	(53.28)	(47.99)	(32.75)	(25.43)	(2.61)
SMB	0.051**	-0.180***	-0.028	0.156***	0.496***	0.445***
	(2.31)	(-5.91)	(-0.58)	(3.61)	(6.96)	(5.18)
HML	0.248***	-0.175***	-0.261***	-0.195***	0.009	-0.239**
	(10.38)	(-4.63)	(-5.34)	(-3.95)	(0.09)	(-2.13)
CMA	-0.080**	-0.006	0.123	0.180	0.245	0.324*
	(-2.31)	(-0.08)	(1.63)	(1.52)	(1.61)	(1.90)
RMW	0.118***	-0.042	-0.101	-0.302***	-0.590***	-0.707***
	(4.27)	(-0.66)	(-1.31)	(-3.19)	(-5.34)	(-5.57)
R^2	0.96	0.90	0.90	0.84	0.79	0.38

The table shows the average excess returns of R&D intensity-sorted portfolios in panel A and results of regressing the portfolio returns on a constant and Fama-French 5 factors in panel B. R&D intensity (RDI) is research and development expenses divided by the market value of equity. Portfolio ”0” consists of firms with missing or no R&D expenditures and the remaining portfolios of firms are sorted by *RDI*. HL is a zero-cost portfolio with a long position in the highest RDI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios each year at the end of June. All portfolios are value-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, but the estimates of the average excess returns and constants are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The sample period is 1975-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.5: R&D-Intensity sorts and performance, Fama-French factors

	PI	PI	RDI	RDI
Constant	5.826***	2.915**	4.737**	1.688
	(2.79)	(1.97)	(2.46)	(1.16)
Mkt-RF	0.140**	0.061	0.129***	0.055
	(2.45)	(1.25)	(2.61)	(1.31)
SMB	0.447***	0.174**	0.445***	0.211***
	(4.11)	(2.26)	(5.18)	(3.29)
HML	-0.355***	-0.208**	-0.239**	-0.053
	(-3.29)	(-2.45)	(-2.13)	(-0.56)
CMA	0.412**	0.212	0.324*	0.109
	(2.15)	(1.49)	(1.90)	(0.83)
RMW	-0.644***	-0.209*	-0.707***	-0.370***
	(-3.70)	(-1.84)	(-5.57)	(-5.08)
RDI		0.614***		
		(10.82)		
PI				0.523***
				(14.16)
R^2	0.34	0.56	0.38	0.58

The table shows the results of regressing PI- and RDI-sorted zero-cost portfolios onto the Fama-French 5 factors as well as the PI and RDI portfolios. PI is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in PI portfolio ”0”. RDI is a zero-cost portfolio with a long position in the highest RDI portfolio and a short position in RDI portfolio ”0”. More details can be found in the descriptions of Table 4.4 and Table 4.5. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The sample period is 1975-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.6: Patent Intensity and R&D Intensity.

measures of technological innovation, capture similar variations in risk and expected returns. Of less interest to us is a “horse race” between the two measures, since in theory, both are important. While our current interest is the similarity between PI and RDI, future research may want to explore their differences, particularly as they should capture different phases of the innovation process. One practical difference between PI and RDI is the considerably longer sample period permitted by patent intensity. Standardized R&D data begins only in 1975, whereas our current study calculates patent intensity for the entire 95 years of available CRSP data. Since reliable patent data goes back even further, until 1790, the only limitation preventing further historical analysis of patent intensity is comprehensive linking to stock return data. A final advantage of the patent data is the lack of ambiguity about the definition of portfolio "0" for PI sorts, as non-patenting firms can be clearly identified from the data. The close resemblance of the RDI portfolio "0" and the PI portfolio "0" serves as a robustness check for the treatment of missing values in the R&D data.

4.3.3 Pricing patent intensity with q-factors

Hou et al. (2015) develop their original q-factor model motivated by the first-order conditions of the static optimization problem of a profit-maximizing firm, suggesting investment and profitability as characteristics related to firm returns.[13] Their q-factor model has four factors, with market and size in addition to investment and profitability.

The q-factor model is sometimes presented as in competition with the five-factor model of Fama and French (2015),[14] but for our purposes, the similarities between the two models are more relevant. Fama and French (2015) also have market, size, investment, and profitability factors, and acknowledge that their value factor is redundant after accounting for the first four factors. Further, the characteristics for size and investment are identical in both approaches. If the value factor is removed, the remaining differences between the two approaches relate to how profitability is defined, and the sorting procedures used for combining factors.[15] Like HXZ, FF have consistently emphasized using simple economic theory to discipline the factors, favoring a static return decomposition (see for example equation 3 in Fama and French (2015) and equation 2 in Fama and French (1995)). While the q-factor model and the FF5 model may certainly have meaningful empirical differences in specific cases, the economic motivation and content of the models are similar, and we expect them to present a consistent overall picture of technological innovators.

A much more important distinction is the $q5$ model of Hou et al. (2021), which adds an expected growth factor. Expected growth fits into the paradigm of appearing in the first-order conditions of an optimizing firm, once extended to a multiperiod setting (see HMXZ equation 1). One can also see that growth matters in the accounting identity of Fama and French (2015), allowing for variation in future quantities (see their equation 3). Technological innovation should naturally be expected to load on expected growth. Innovation creates new products or reduces costs, raising the marginal product of future investments in traditional assets, and adding to expected growth. Correspondingly, HMXZ demonstrate the importance of expected growth for R&D sorted portfolios.

[13] See their equation 4. Earlier literature documents the anomalies related to investment (Titman et al., 2004; Cohen et al., 2013) and profitability (e.g., Novy-Marx (2013b)).

[14] See for example Hou et al. (2020b).

[15] Fama and French (2015) define profitability as operating profitability scaled by annually updated book equity while Hou et al. (2015) using earnings before extraordinary items scaled by quarterly updated book equity. FF use bivariate sorts on size and profitability and size and investment to form those factors, while HXZ use a trivariate sort of all three characteristics.

We show that the expected growth factor also plays an essential role in pricing patent intensity portfolios. This complements the findings of HMXZ by using a different but related measure of technological innovation. Further, our sample is nine years longer, limited only by the availability of q-factors before 1967. Finally, our methodology uses a broader cross-section of firms, including the portfolio zero of non-innovators.

We first apply the original q-factor model with four factors. Panel A in Table 4.7 shows that this model leads to similar or even stronger mispricing across the portfolios than the FF5 model. The alphas increase monotonically from -2.03% in portfolio zero to 6.79% in portfolio 4, generating abnormal return of 8.82% for the HL portfolio. In unreported results, we confirm that the stronger mispricing result relative to the FF5 model is not driven by the slightly later start of the q-factor data in 1967.

The q-factor loadings of the PI-sorted portfolios closely resemble the loadings on the five Fama-French factors discussed in the previous section. In particular, the loadings on the profitability factor (ROE) decrease almost monotonically across the portfolios from slightly positive but insignificant value for non-patenting firms to significantly negative value of -0.48 for high-PI firms. This lowers the q-factor model implied expected returns for high-PI portfolios, further adding to the already higher excess returns of these portfolios. Although the q-factor model is based on an appealing investment asset pricing framework, its empirical factors share some key characteristics of the FF5 factors, and hence lead to a similar amplification of the mispricing of the PI-sorted portfolios.

Panel B shows q5-regressions, which include the expected growth factor (EG). The loadings show a strong relationship between patenting intensity and expected growth. Non-patenting firms have a negative loading of -0.18 on EG, which monotonically increases with PI to 0.64 for high-patent intensity firms, generating a loading spread of 0.82 in the long-short portfolio. Including EG further amplifies the negative loadings on the investment and profitability factors, which decrease to -0.42 and -0.79 from -0.27 and -0.52, respectively. The inclusion of the EG factor is crucial to explain the returns of the PI-sorted portfolio. While q5-factor alphas are still monotonically increasing, resulting in a long-short alpha of 2.26%, the remaining long-short alpha is statistically indistinguishable from zero. The results show that the q5 model is able to price technological innovators, in particular relative to non-innovating firms.

In unreported robustness checks, we find that the statistical significance of the q5 results is sensitive to the exact specification of patenting activity. For example, when measuring patent intensity as the number of patents over the last 36 months (instead of 12 months in the main specification) divided by the firm's market capitalization, the high-PI portfolio still earns a statistically significant q5-alpha, resulting in a significant alpha of the HL portfolio as well. Nonetheless, what remains robust across all checks is a very strong sort on the expected growth loading and a large reduction

	0	1	2	3	4	HL
Panel A. Q4-factors 1967-2021						
Constant	-2.04***	-0.04	2.83***	3.5***	6.79***	8.82***
	(-3.27)	(-0.08)	(3.75)	(2.87)	(3.69)	(3.96)
MKT	1.0***	0.95***	0.99***	1.05***	1.08***	0.08
	(50.36)	(86.86)	(53.28)	(28.13)	(19.58)	(1.11)
ME	0.14***	-0.19***	-0.08**	0.2**	0.55***	0.41**
	(2.68)	(-11.85)	(-2.07)	(2.56)	(4.44)	(2.43)
IA	0.22***	-0.06**	-0.05	-0.06	-0.05	-0.27*
	(4.52)	(-2.43)	(-1.25)	(-0.85)	(-0.39)	(-1.76)
ROE	0.04	0.07***	-0.15***	-0.29***	-0.48***	-0.52***
	(1.33)	(3.14)	(-3.6)	(-4.64)	(-5.02)	(-4.64)
R^2	0.95	0.96	0.9	0.86	0.78	0.28
Panel B. Q5-factors 1967-2021						
Constant	-0.59	0.18	0.57	1.18	1.67	2.26
	(-1.06)	(0.39)	(0.76)	(0.98)	(1.05)	(1.24)
MKT	0.98***	0.95***	1.02***	1.08***	1.15***	0.17***
	(55.85)	(86.26)	(60.46)	(29.18)	(23.18)	(2.85)
ME	0.12**	-0.19***	-0.05	0.23***	0.61***	0.49***
	(2.36)	(-12.25)	(-1.44)	(2.99)	(5.12)	(2.98)
IA	0.25***	-0.06**	-0.11***	-0.12*	-0.17	-0.42***
	(5.13)	(-2.15)	(-2.59)	(-1.7)	(-1.39)	(-2.8)
ROE	0.1***	0.08***	-0.25***	-0.38***	-0.69***	-0.79***
	(2.7)	(3.31)	(-5.57)	(-5.35)	(-6.84)	(-6.37)
EG	-0.18***	-0.03	0.28***	0.29***	0.64***	0.82***
	(-4.88)	(-0.9)	(5.54)	(3.8)	(5.92)	(6.52)
R^2	0.95	0.96	0.91	0.86	0.79	0.35

The table shows the results of regressing the PI-sorted portfolio returns on a constant and the four Q-factors (Hou et al., 2015), i.e., market (MKT), size (ME), investment (IA), and profitability (ROE), in panel A, and additionally on fifth Q-factor (Hou et al., 2021), i.e., expected growth (EG), in panel B. Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios each year at the end of June. All portfolios are value-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, and the estimates of the constants are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the Q-factors, i.e., 1967-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.7: Patent Intensity and q-factors

in mispricing. These are the key economic findings that we focus on.

4.3.4 Build-up or resolution?

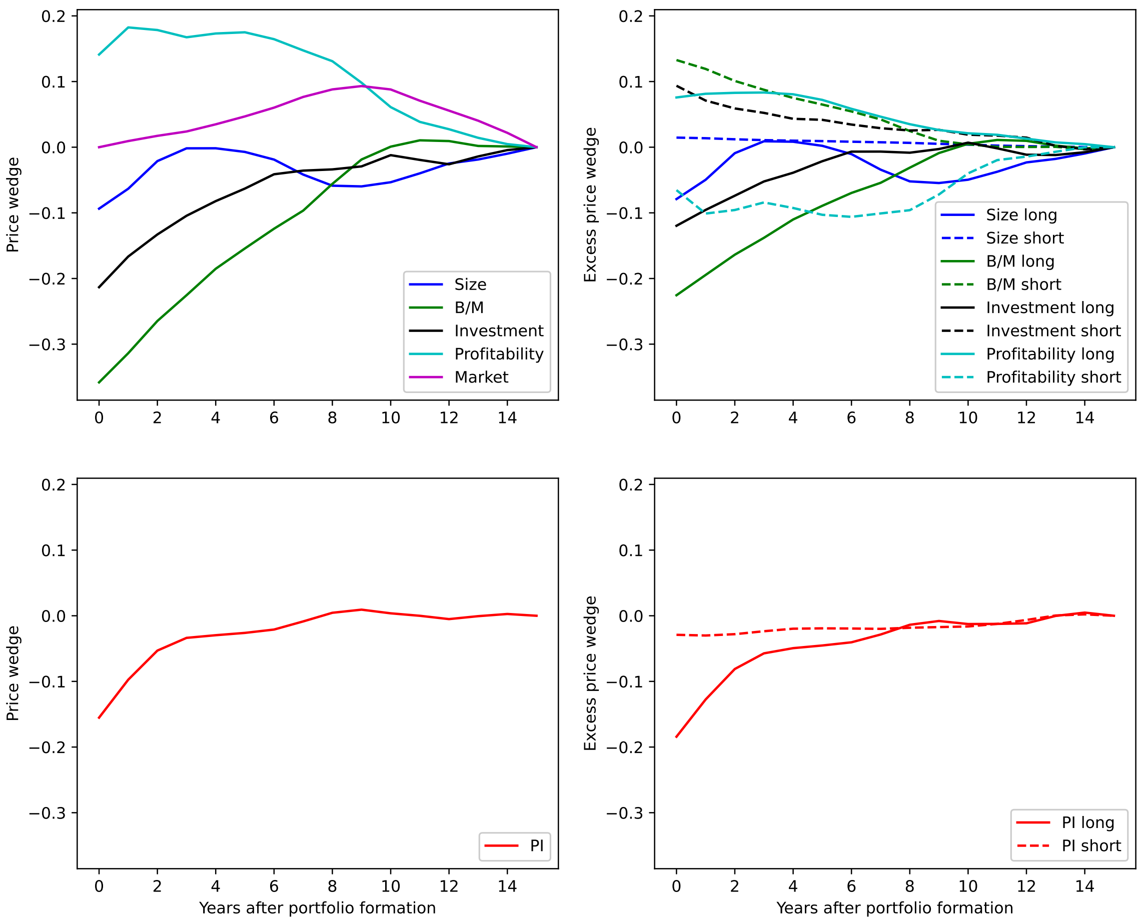
A different lens through which to understand the performance of technological innovators is the methodology of van Binsbergen et al. (2021), which proposes to determine whether an anomaly is due to "build-up" or "resolution" of misvaluation. They generate an empirical pricing kernel by assuming that the market portfolio is priced correctly on average over the sample period, given realized cash flows (dividends) over a fifteen-year period and the terminal value of the portfolio in year fifteen. Other portfolios, such as the market at other horizons or any anomaly portfolio at any horizon, can be valued using this pricing kernel. Assets are therefore priced by their covariation with realized market returns, as in the CAPM. We apply this methodology to our patent-intensity portfolios.

Starting in 1963, we estimate the fair market value of anomaly portfolios, including PI-portfolios, using the van Binsbergen et al. (2021) dividend discount model and CAPM-SDF. For greater comparability with their results, we form our last portfolios in 2002 (final cash flows in 2017). Portfolios are therefore formed in June of every year from 1963 to 2002. The price wedge of a portfolio is the difference between the actual price of the portfolio and the imputed fair market value from the model. In addition to the price wedge at the time of portfolio formation, we track the portfolios through time until 15 years after portfolio formation. Importantly, we track the same group of stocks throughout the 15 years and keep the endpoint constant, forcing the price wedge to be equal to zero after 15 years. We carry out this methodology for not only the PI-portfolios but also the market and anomalies related to size, value, investment, and profitability.

Fig. 4.4 shows estimated price wedges. The top left panel shows the benchmark market portfolio, and long-short portfolios formed on size, value, investment, and profitability. This reveals an important consideration in interpreting the reported price wedges: The market itself is "misvalued" in the years after portfolio formation as it ages. We point this out not to critique the methodology, but to make clear that the pattern observed in the market is the benchmark by which we may want to evaluate other portfolios.[16] The long-short portfolios in the top left corner should not be as strongly affected by this benchmark issue, since it affects both the long and short sides. Consistent with the results of van Binsbergen et al. (2021), the profitability anomaly is a "build-up" anomaly and the other anomalies considered are "resolution" or reduction of existing mispricing.

The top right panel of Fig. 4.4 shows price wedges for the long and short sides separately of

[16]The apparent misvaluation of the market at intermediate horizons could be due to autocorrelations in market returns, or to dropping years of data at the sample beginning in the aged portfolios. For example, the one-year aged portfolio drops from the valuation of the market all of the 1963 data.



The figure shows price wedge dynamics for portfolios sorted on size, book-to-market, investment, and profitability in the top row and portfolios sorted on PI in the bottom row. Price wedges are calculated as the difference between observed and the fair market value suggested by a 15-year dividend discount model using CAPM SDF as suggested by van Binsbergen et al. (2021). The top-left panel plots the price wedge for a long-short portfolio, where the long side is the quartile portfolio with the highest (lowest) value of b/m or profitability (size or investment) and the short side is the quartile portfolio with the lowest (highest) value. *Market* is the estimated price wedge of the market portfolio. The top-right panel plots the price wedges of the individual legs of the aforementioned long-short portfolios. The bottom-left panel plots the price wedge of a portfolio that goes long high PI firms (portfolio ”4”) and short low PI firms (portfolio ”0”). The bottom-right panel shows the wedges of the two portfolios separately.

Figure 4.4: Price wedge dynamics

each of the traditional anomalies. To avoid the benchmark issue shown for the market portfolio, we display price-wedge differences, the difference between the price wedge of each portfolio and the price wedge of the market portfolio (if we did not subtract the price wedge of the market portfolio, all long-only portfolios would have this as a common component of their price wedges). Undervaluation appears to play a modestly more important role than overvaluation. We also see differences in the speed of misvaluation resolution. For example, small stocks have a small undervaluation wedge that dissipates quickly.

The bottom two panels of Fig. 4.4 show price wedge dynamics of the PI-sorted portfolios, with the long-short portfolio in the left-hand panel and the price wedge differences (relative to market) of the long and short sides separately on the right-hand side. According to the benchmark model, the long-short portfolio is initially undervalued by a little less than twenty percent, with all of this coming from undervaluation of patent-intensive firms.

These results help to interpret the CAPM results shown in Table 4. According to the CAPM, non-innovators (the short side of the PI long-short portfolio) are not mispriced, and the price wedge shows no long-run mispricing either. On the other hand, patent-intensive firms earn positive abnormal returns, and the bottom right-hand panel of Fig. 4.4 says that this should be interpreted as undervaluation that takes several years to resolve. The results thus conform well with early discussions in the literature of undervaluation of technological innovation by investors, perhaps because of short-sightedness or misunderstanding the value of innovation (Hall, 1993; Hall and Hall, 1993).

A natural question to follow this analysis is why adding additional "standard" factors to the CAPM, such as investment and profitability, worsens the mispricing of technological innovators (Table 4.4, Panel D)? Further, is this additional mispricing short-lived or long-lived? We turn to these questions in the next section.

4.4 (Mis)pricing innovation

We further explore the dynamics of mispricing in patent-intensity portfolios by considering factor-model regressions on "aged" portfolios, formed in year zero, and followed for up to ten years following formation. Consistent with previous results, CAPM and FF3 abnormal performance resolves within two to three years of portfolio formation. Adding investment and profitability factors, in FF5 and $q4$, dramatically worsens the picture. Abnormal returns on the long-short PI portfolio remarkably remain significantly positive for a full decade after portfolio formation.

The expected growth factor of the $q5$ model again largely resolves the problem. For almost all PI-sorted portfolios and horizons, abnormal performance becomes insignificant. Loading dynamics explain the lives of innovators. Expected growth starts very high and remains high for a full decade while gradually declining. Investment in traditional assets becomes more aggressive in the years immediately following formation and levels off. Profitability begins very weak and gradually improves throughout the decade. Removing any one of these priced sources of fundamentals causes persistent mispricing of innovators.

To further understand the role of investment and profitability, we carry out characteristic sorts *within* portfolios of innovators and non-innovators to see if the characteristics earn similar spreads. Market, size, and value characteristics largely earn similar spreads among innovators and non-innovators, both in raw returns and controlling for standard factors. But investment and profitability are different. Among non-innovators, high investment is associated with the familiar lower returns. Among innovators, the spread is smaller and not statistically significant. Controlling for standard factors, non-innovators are priced well, but innovators with high investment rates have positive abnormal returns. Profitability sorts produce a more extreme result. Among non-innovators, profitability associates with the familiar higher return. But among innovators, the return spread has the opposite sign, though not statistically significant. As a consequence, applying standard factor models, non-innovators appear well-priced, but innovators earn a significant positive alpha for weak profitability. In sum, innovators that load on investment and profitability do not earn the same return premium as the overall population, explaining the alphas earned by exposure to these factors among innovators.

We conclude by examining the ability of the expected growth factor of the $q5$ model to resolve the challenging mispricing of characteristic sorts within innovators, and discuss implications for future research.

4.4.1 Risk and alpha dynamics of aged portfolios

We examine risk and alpha dynamics of patent-intensity sorted portfolios for a decade following the initial sort date. At the end of June of year t we use the PI sort from year $t-K$ to form value-weighted portfolios, for lags $K = 0, 1, \ldots, 10$. The sorts do not depend on time-t information, and any stocks from the $t-K$ sort that are no longer present at date t are simply omitted from the aged portfolio. The value weights depend on values at the end of June of year t. The K-aged portfolio returns are identical to the returns one would receive if forming the portfolios at year $t-K$, rebalancing each year to current value weights based on the stocks remaining from the original portfolio sort, and reinvesting any dividends or delisting returns at the same value weights. In other words, we study portfolios of firms that were classified as high-PI or non-patenting K years ago. The analysis reveals the evolution of risk and performance of the initially sorted portfolios over time.

Table 4.8 shows FF5 alpha dynamics of the aged portfolios in the 1963-2021 sample period.

The results are striking. FF5 abnormal performance for non-innovators is significantly negative for a full eleven years after formation (cohorts 0 to 10), and the high-PI portfolio remains significantly positive for a full ten years. The long-short portfolio alpha is highly statistically significant exceeding 5% annually in the 10th year after formation (cohort 9). The persistence of performance is remarkable.

Table 4.9 shows long-short returns and alphas for CAPM, FF3, and FF5, for the full sample and post-1963 sample. For the CAPM and FF3 models, positive abnormal returns remain statistically significant for only two to three years. The addition of the investment and profitability factors in the FF5 portfolios not only makes abnormal performance larger but also substantially more persistent. As discussed by van Binsbergen and Opp (2019), persistence in abnormal performance, or significant inaccuracy in costs-of-capital over long periods of time, can imply highly inefficient real investment. If the FF5 model accurately captures the market-required return on equity capital, technological innovators face too-high costs of capital for long horizons, and are therefore likely to significantly underinvest. Meanwhile, non-innovator costs-of-capital would be too low, implying overinvestment. Table C.4 and Table C.5 in the appendix show similar results respectively for FF6 alphas (adds momentum) and $q4$ alphas.

Table 4.10 shows that the expected growth factor of the $q5$ model again remarkably resolves these difficulties for nearly all portfolios and horizons. To understand the role played by expected growth, Fig. 4.5 shows the dynamics of factor loadings in the $q5$ model (Fig. C.1 and Fig. C.2 in the appendix show similar loadings for FF5 and $q4$ models). Table 4.7, Panel B previously showed a very high contemporaneous loading of the long-short PI portfolio on expected growth, but what does such a high level of expected growth imply for the risk dynamics of technological innovator loadings?

The factor loading dynamics reveal a compelling economic story. First, we consider expected growth itself. The initial spread is very strong and monotonic, with the high-PI loading exceeding 0.6, the non-innovator loading approaching -0.2, and the net long-short loading exceeding 0.8. Over time, we should always anticipate loadings with a strong initial sort to mean-revert. The growth loadings mostly do so, but with a twist. In particular, the four innovator portfolio loadings appear to mean revert to a common mean, and all are in-between 0.1 and 0.2 in the 10th year, while the non-innovator loading stays negative and statistically significant throughout the decade. The long-short growth loading is 0.31 with a t-statistic exceeding 2 in the 10th year. Innovator growth and non-innovator growth appear to revert to different means, and innovator growth loadings are persistently higher.

The loadings on investment also show a strong distinction between innovators and non-innovators. The non-innovator investment loading is in the range of 0.25 to 0.3 (conservative) and highly statis-

Horizon (years)	0	1	2	3	4	HL
0	-1.76***	0.13	2.11***	2.02**	4.94***	6.71***
	(-3.75)	(0.3)	(3.03)	(1.97)	(2.96)	(3.54)
1	-1.78***	-0.13	2.57***	1.38	6.24***	8.03***
	(-3.9)	(-0.31)	(3.8)	(1.48)	(4.01)	(4.6)
2	-1.75***	0.0	1.55**	1.0	4.88***	6.63***
	(-3.71)	(0.01)	(2.5)	(1.2)	(3.19)	(3.83)
3	-2.0***	-0.13	1.53**	1.39	2.5*	4.5***
	(-4.26)	(-0.3)	(2.34)	(1.6)	(1.83)	(2.87)
4	-1.9***	-0.08	0.95	1.74*	1.92	3.82**
	(-4.07)	(-0.18)	(1.57)	(1.93)	(1.49)	(2.51)
5	-1.84***	0.18	0.44	1.73*	2.22*	4.07***
	(-3.99)	(0.38)	(0.72)	(1.94)	(1.71)	(2.71)
6	-1.82***	0.27	0.58	0.88	3.31**	5.13***
	(-3.98)	(0.59)	(1.02)	(0.94)	(2.33)	(3.15)
7	-1.55***	0.13	0.33	1.41*	2.97**	4.52***
	(-3.47)	(0.31)	(0.58)	(1.65)	(2.09)	(2.79)
8	-1.66***	-0.03	0.92	0.46	2.7*	4.36***
	(-3.7)	(-0.08)	(1.5)	(0.48)	(1.9)	(2.68)
9	-1.71***	0.27	0.35	0.76	3.55**	5.26***
	(-3.74)	(0.6)	(0.6)	(0.87)	(2.48)	(3.18)
10	-1.43***	0.24	-0.67	2.04**	1.44	2.86*
	(-3.09)	(0.5)	(-1.12)	(2.11)	(0.99)	(1.69)

The table shows the abnormal returns (alphas) relative to five-factor model (Fama and French 2015) of PI-sorted portfolios for holding period of one-year at different investment horizons (indicated in rows). Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios at the end of June K years prior to the beginning of the holding period in July of year t. The holding period lasts for one year from July (end of June) in year t to the end of June in year $t+1$. Each portfolio consists of the stocks assigned to the portfolio K years ago that are still active as of the beginning of the holding period, i.e., end of June in year t. Portfolios are value-weighted with weights as of the beginning of the holding period. The underlying portfolio returns are at monthly frequency, but the estimates of the alphas are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the FF5-factors, i.e., 1963-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.8: Aged Patent Intensity portfolios, FF5 alpha dynamics, 1963-2021

Horizon (years)	Panel A. 1926-2021 Excess return	CAPM alpha	FF3 alpha	Panel B. 1963-2021 Excess return	CAPM alpha	FF3 alpha	FF5 alpha
0	7.187***	3.914***	3.392**	6.968***	5.119***	4.99***	6.709***
	(4.423)	(2.852)	(2.505)	(3.431)	(2.605)	(2.604)	(3.543)
1	7.002***	3.88***	3.649***	7.213***	5.279**	5.849***	8.027***
	(4.362)	(2.651)	(2.648)	(3.507)	(2.535)	(3.202)	(4.604)
2	4.726***	2.056	2.201	4.736**	2.862	4.021**	6.628***
	(3.074)	(1.419)	(1.631)	(2.381)	(1.373)	(2.299)	(3.83)
3	3.169**	0.474	0.58	2.882	0.803	1.827	4.504***
	(2.05)	(0.337)	(0.436)	(1.495)	(0.402)	(1.039)	(2.874)
4	2.772*	0.156	0.221	2.7	0.736	1.638	3.822**
	(1.941)	(0.118)	(0.175)	(1.499)	(0.4)	(1.004)	(2.514)
5	3.976***	0.935	0.931	3.425*	1.321	2.369	4.067***
	(2.754)	(0.73)	(0.751)	(1.942)	(0.758)	(1.431)	(2.711)
6	5.019***	1.636	2.063	4.469**	2.218	3.406**	5.13***
	(3.356)	(1.197)	(1.584)	(2.367)	(1.204)	(1.997)	(3.149)
7	3.42**	0.484	1.309	3.483*	1.296	2.725	4.518***
	(2.381)	(0.345)	(1.004)	(1.823)	(0.674)	(1.61)	(2.79)
8	3.67**	0.731	1.847	3.324*	1.036	2.495	4.355***
	(2.45)	(0.501)	(1.373)	(1.677)	(0.522)	(1.437)	(2.681)
9	3.329**	0.726	2.137	3.823*	1.663	3.18*	5.26***
	(2.242)	(0.487)	(1.552)	(1.89)	(0.822)	(1.828)	(3.181)
10	1.8	-0.784	0.51	2.313	0.027	1.42	2.865*
	(1.24)	(-0.545)	(0.377)	(1.178)	(0.014)	(0.817)	(1.693)

The table shows the excess and abnormal return (indicated in columns) on PI-sorted long-short portfolios for holding period of one-year at different investment horizons (indicated in rows). The PI-sorted long-short portfolio consists of a long position in high-PI firms and a short position in non-patenting firms. In panel A, stocks are sorted into three portfolios (non-patenting, low-PI and high-PI), and in panel B into five portfolios (non-patenting, and the remaining patenting stocks into four portfolios by PI). Stocks are sorted into portfolios at the end of June K years prior to the beginning of the holding period in July of year t. The holding period lasts for one year from July (end of June) in year t to the end of June in year $t+1$. Each portfolio consists of the stocks assigned to the portfolio K years ago that are still active as of the beginning of the holding period, i.e., end of June in year t. Portfolios are value-weighted with weights as of the beginning of the holding period. Excess return is average return of the long-short portfolio in excess of risk-free rate. CAPM alpha, FF3 alpha, and FF5 alpha indicate abnormal return relative to market model, Fama and French 1993, and Fama and French 2015, respectively. The underlying portfolio returns are at monthly frequency, but the estimates of the average excess returns and alphas are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.9: Aged Patent Intensity long-short portfolios, alpha dynamics

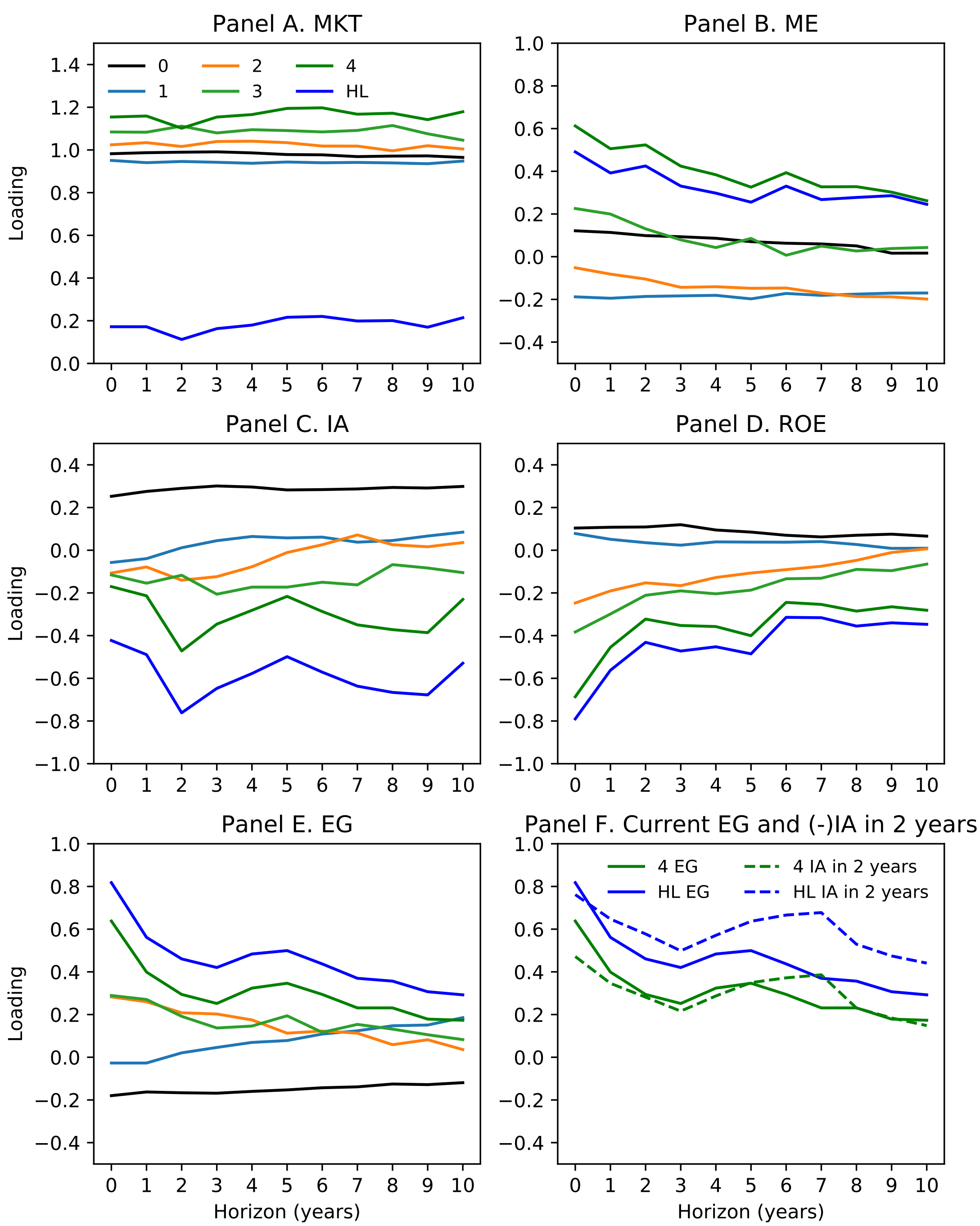


Figure 4.5: Aged Patent Intensity portfolios, q5-factor loading dynamics, 1967-2021

Horizon (years)	0	1	2	3	4	HL
0	-0.59	0.18	0.57	1.18	1.67	2.26
	(-1.06)	(0.39)	(0.76)	(0.98)	(1.05)	(1.24)
1	-0.76	0.27	0.69	0.11	3.57	4.33*
	(-1.31)	(0.54)	(0.86)	(0.1)	(1.56)	(1.65)
2	-0.65	0.01	0.22	0.05	2.9	3.56
	(-1.09)	(0.01)	(0.27)	(0.05)	(1.37)	(1.47)
3	-1.05*	-0.23	0.54	0.91	0.71	1.76
	(-1.85)	(-0.45)	(0.65)	(0.8)	(0.44)	(0.93)
4	-0.91	-0.5	0.15	1.28	-0.42	0.49
	(-1.62)	(-0.95)	(0.2)	(1.15)	(-0.27)	(0.27)
5	-0.94*	-0.3	0.22	0.76	-0.28	0.66
	(-1.68)	(-0.54)	(0.27)	(0.69)	(-0.19)	(0.38)
6	-0.95*	-0.35	0.12	0.51	0.45	1.4
	(-1.68)	(-0.62)	(0.17)	(0.46)	(0.27)	(0.71)
7	-0.65	-0.78	0.02	0.66	0.98	1.63
	(-1.17)	(-1.49)	(0.03)	(0.65)	(0.58)	(0.84)
8	-1.09**	-0.87	0.95	-0.47	0.88	1.97
	(-1.98)	(-1.4)	(1.2)	(-0.45)	(0.5)	(0.97)
9	-1.08*	-0.49	0.05	-0.01	2.08	3.16
	(-1.87)	(-0.84)	(0.07)	(-0.01)	(1.17)	(1.53)
10	-0.86	-0.96	-0.5	1.36	0.11	0.97
	(-1.41)	(-1.55)	(-0.63)	(1.16)	(0.06)	(0.46)

The table shows the abnormal returns (alphas) on PI-sorted portfolios for holding period of one-year at different investment horizons (indicated in rows) relative to the Q-factor model (Hou et al., 2021). Portfolio 0 consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio 0. Stocks are sorted into portfolios at the end of June K years prior to the beginning of the holding period in July of year t. The holding period lasts for one year from July (end of June) in year t to the end of June in year $t+1$. Each portfolio consists of the stocks assigned to the portfolio K years ago that are still active as of the beginning of the holding period, i.e., end of June in year t. Portfolios are value-weighted with weights as of the beginning of the holding period. The underlying portfolio returns are at monthly frequency, but the estimates of the alphas are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the Q-factors, i.e., 1967-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.10: Aged Patent Intensity portfolios, q-factor alpha dynamics, 1963-2021

tically significant throughout the decade. Innovator loadings are initially negative (aggressive) and bunched (-0.06 to -0.17), but then diverge. Low-intensity innovators become more conservative in their investment loadings and high-intensity innovators become more aggressive. High-intensity innovators particularly shift toward aggressive investment in the two years following portfolio formation. To explore the relation between expected growth and investment further, Panel F (bottom right) plots growth loadings and two-year forward investment loadings on the same axes. The overlap is remarkably strong. Expected growth loadings predict future investment loadings for technological innovators.

The final piece of the economic story is profitability loadings. Once again the initial sort is strong and monotonic, with non-innovators loading slightly positively on profitability (0.1, t=2.7) and high-intensity innovators loading negatively (-0.69, t=-6.8). Over time mean-reversion occurs, but slowly and mostly among the most innovation-intense firms. In the 10th year, the loading sort is still monotonic, with non-innovators still loading positively (0.08, t=1.9) and high-intensity innovators still loading negatively (-0.26, t=2.7). Over the ten-year period, non-innovator profitability very modestly weakens (year 10 minus year 0 profitability loading equals -0.04, t=-2.36, Table C.9, Panel D in the appendix). Meanwhile, the most intense innovators move strongly towards more robust profitability (year 10 minus year 0 profitability loading equals 0.4, t=4.9).

These three elements, growth, investment, and profitability, drive a compelling economic story. High-intensity innovators develop growth options, which they take advantage of through increasingly heavy investment, gradually leading to improved profitability. All three factors earn strong premia, and all are needed to explain the complex risk and return dynamics of innovative firms.

Though not as central to the economic story, the dynamics of size loadings are also interesting. Naturally, we expect *ex-ante* that size loadings should decrease, as the firms in the aged portfolios are not replaced by new firms. Most of the portfolios follow a pattern of gradual decrease in size loadings, but size loadings drop most dramatically for the most innovation-intensive, consistent with these firms growing fastest.

4.4.2 Investment and profitability

We finally seek to further understand why investment and profitability factors can misprice technological innovators when not controlling for expected growth. We also more generally explore the question of differences in pricing between innovators and non-innovators. Our approach is to sort *within* the groups of all non-innovators ($PI = 0$, portfolio 0) and all innovators ($PI > 0$, portfolios 1-4) on the FF5 characteristics: market, size, B/M, profitability, and investment. For simplicity, the long-short portfolios are always long the quintile with the highest value of the sorting variable and short the quintile with the lowest value, irrespective of which side earns the higher return traditionally. We ask whether the characteristics earn similar return spreads within the groups of innovators and non-innovators, and compare alphas after controlling for the FF5 factors.

Table 4.11 shows results. For beta, size, and B/M the return spreads and alphas are largely unremarkable. The value spreads are positive and significant among both innovators and non-innovators, but their difference is not, and none of the alphas is significant controlling for standard factors. The raw size spread is larger for innovators than non-innovators, but the alpha difference is insignificant controlling for standard factors. The raw beta spreads are insignificantly different from zero, as are the within-group alphas, but the alpha is mildly larger for innovators than non-innovators (beta earns more of a premium for innovators than non-innovators). These results do not appear central to explaining the pricing of innovative versus non-innovative firms.

The results for investment and profitability are more noteworthy. The raw return spread for non-innovators shows the familiar negative sign and is statistically significant, whereas the return spread for innovators is negative, but of lower magnitude and not significant. The difference in spreads is not significant. Controlling for FF5 factors, however, the difference in alpha becomes significantly positive, with a magnitude of 4.5% p.a. (t=2.36) driven by a very negative loading on investment (innovators have a much wider spread in investment loadings than non-innovators). The difference for profitability is even stronger. In raw returns, among non-innovators the most profitable quintile earns the familiar higher return than the least profitable quintile, with a return spread of 6% p.a. (t=2.34). To the contrary, among non-innovators the return spread is *negative* (-3.5% p.a.) but not significant. The return spread difference is very large, -9.6% p.a. (t=-3.4). Controlling for the FF5 factors has predictable results. The alpha for non-innovators is statistically indistinguishable from zero, but the alpha for innovators is -6.2% p.a. (t=-2.7). The alpha difference is economically meaningful at -7.8% p.a., and statistically significant.

The explanation for mispricing when using the FF5 model for innovative firms is now clear. In the aggregate data, the return spreads earned for investment and profitability are driven primarily by non-innovators. Innovators have strong variation in these characteristics, but the return spreads are weaker or even opposite to the overall data.

Why does the $q5$ model help to price portfolios of innovators, and can it further solve the challenging problem of characteristic sort mispricing *within* groups of innovators and non-innovators? Table 4.12 sheds light on these questions, showing mixed results. For investment sorts, the alpha difference between innovators and non-innovators falls by more than 50% to approximately 2% p.a. and a t-statistic less than one. The improvement in pricing is driven by a strong positive loading on expected growth (0.36, raising the benchmark required return) that partially compensates for the large negative investment loading (-0.74, decreasing the benchmark required return). This is the classic omitted variables problem. Heavy investors who are innovators are expected to grow faster

		Ex. Ret.	Fama-French 2015						
		Constant	Constant	Mkt-RF	SMB	HML	CMA	RMW	R^2
Beta	Non-Inno	-1.237	-1.750	0.619***	-0.163***	0.140*	-0.762***	-0.372***	0.40
		(-0.49)	(-0.85)	(14.61)	(-2.68)	(1.72)	(-6.27)	(-4.42)	
	Inno	0.093	2.734	0.435***	-0.065	-0.233**	-0.751***	-0.645***	0.33
		(0.03)	(1.10)	(8.48)	(-0.88)	(-2.35)	(-5.10)	(-6.32)	
	Diff	1.330	4.484*	-0.183***	0.098	-0.374***	0.011	-0.273***	0.07
		(0.59)	(1.96)	(-3.87)	(1.45)	(-4.09)	(0.08)	(-2.91)	
Size	Non-Inno	-4.582*	-3.732*	0.225***	-1.148***	-0.320***	0.207*	0.257***	0.42
		(-1.82)	(-1.85)	(5.41)	(-19.25)	(-3.98)	(1.73)	(3.11)	
	Inno	-7.833***	-4.125**	-0.099**	-1.482***	-0.338***	0.070	0.421***	0.62
		(-2.67)	(-2.17)	(-2.52)	(-26.39)	(-4.47)	(0.62)	(5.41)	
	Diff	-3.251*	-0.393	-0.324***	-0.334***	-0.018	-0.137	0.164**	0.25
		(-1.73)	(-0.23)	(-9.15)	(-6.58)	(-0.27)	(-1.35)	(2.33)	
B/M	Non-Inno	3.792**	-0.217	0.004	0.184***	1.009***	0.136**	-0.202***	0.62
		(2.16)	(-0.19)	(0.19)	(5.47)	(22.23)	(2.02)	(-4.34)	
	Inno	5.965***	-0.794	0.157***	0.424***	0.976***	0.370***	-0.020	0.45
		(2.68)	(-0.46)	(4.39)	(8.27)	(14.13)	(3.60)	(-0.28)	
	Diff	2.173	-0.577	0.153***	0.240***	-0.033	0.234*	0.182**	0.05
		(1.10)	(-0.29)	(3.67)	(4.02)	(-0.40)	(1.95)	(2.21)	
Invest	Non-Inno	-4.168***	-0.786	0.007	-0.144***	-0.131***	-0.803***	0.020	0.52
		(-3.51)	(-0.91)	(0.37)	(-5.64)	(-3.81)	(-15.69)	(0.56)	
	Inno	-2.190	3.721**	-0.080**	-0.041	0.037	-1.631***	-0.002	0.48
		(-1.08)	(2.43)	(-2.54)	(-0.90)	(0.61)	(-18.02)	(-0.03)	
	Diff	1.978	4.507**	-0.087**	0.103*	0.168**	-0.827***	-0.022	0.10
		(1.03)	(2.36)	(-2.20)	(1.83)	(2.22)	(-7.33)	(-0.28)	
Profit	Non-Inno	6.048**	1.641	-0.092**	-0.335***	-0.268***	0.440***	1.572***	0.55
		(2.34)	(0.90)	(-2.47)	(-6.27)	(-3.72)	(4.11)	(21.33)	
	Inno	-3.538	-6.208***	-0.197***	-0.387***	0.243***	-0.284**	1.498***	0.47
		(-1.19)	(-2.71)	(-4.21)	(-5.74)	(2.68)	(-2.10)	(16.12)	
	Diff	-9.586***	-7.849***	-0.105*	-0.052	0.511***	-0.724***	-0.075	0.04
		(-3.40)	(-2.69)	(-1.77)	(-0.61)	-4.43	(-4.22)	(-0.63)	

The table shows the average excess returns of innovative and non-innovative firms sorted on common firm characteristics as well as the results of regressing the portfolio returns on a constant and Fama-French 5 factors. Stocks are labeled as innovators and non-innovators at the end of June in each year and then sorted into five portfolios within the two groups. Innovative firms are firms that have at least three patents over the last three years and one patent over the last year at the time of portfolio formation. The table shows the returns of a long-short portfolio that goes long the highest quintile and short the lowest quintile. All portfolios are value-weighted and rebalanced annually. The five firm characteristics (beta, size, book-to-market equity ratio, investments, and profitability), shown in the first column of the table, follow the definitions from Ken French's website. The underlying portfolio returns are at monthly frequency, but constants are expressed in annualized percent. The time period of the sample is 1963-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table 4.11: Characteristics sorts for innovative vs. non-innovative firms

than heavy investors who are non-innovators, and failing to account for this correlation causes mispricing. Turning to profitability sorts, the $q5$ model eliminates the statistical significance of the difference in alpha between innovators and non-innovators, but leaves a strong and statistically significant negative alpha (-7.2%, t=-2.4) for the sort within innovators. In other words, innovators still earn surprisingly low returns for profitability (according to the $q5$ model), even after accounting for the expected growth factor.

Overall, these results shed further light on differences in pricing for innovators and non-innovators. The raw investment and profitability anomalies are stronger in non-innovators than innovators, and innovators show abnormal performance for exposure to these factors using standard FF5 or $q4$ models. The expected growth factor of the $q5$ model realigns pricing of investment because heavy investors who are innovators also tend to have high growth loadings. The pricing of profitability sorts among innovators remains a challenge even for the $q5$ model. We note that the construction of the expected growth factor in the $q5$ model specifically targets investment growth. But fundamental valuation suggests that different types of growth can be relevant, for example, not just investment growth but also profitability growth or revenue growth. We leave these issues for future research.

4.5 Conclusion

Over the past century, approximately a quarter of US-publicly listed firms could be classified as technological innovators by their patenting activity. Since the 1930s, innovators accounted for more than half of the total market capitalization at any point in time. Despite being long proposed as a key driver of economic growth, leading factor models only implicitly take into account technological innovation. Our paper proposes a simple patent-based measure of innovation intensity that allows us to study the role of technological innovation for stock returns.

Technological innovators earn higher returns than non-innovators and do not incur the same punishment for high capital investment and low profitability as non-innovators. In particular, a portfolio of firms with high patenting intensity earns significant abnormal returns for a full decade after portfolio formation, according to standard pricing models. We unite our findings with the recent literature on the role of expected growth in stock returns (Hou et al., 2021). Over time, firms with high patenting intensity invest more in physical capital and gradually improve their profitability as they age. An expected growth factor is crucial to explain the returns of innovating firms.

Our study highlights strongly predictable patterns in the risk dynamics of innovative firms. The results suggest more formally linking theory to the evolution of firm risk, providing stronger tests of pricing models. Since our measure does not rely on accounting data, empirical studies can use long samples, even beyond the nearly full century of data that we study. This is particularly important

		Ex. Ret.	Hou, Mo, Xue, and Zhang 2021						
		Constant	Constant	MKT	ME	IA	ROE	EG	R^2
Beta	Non-Inno	-1.941	0.003	0.651***	-0.200***	-0.248**	-0.036	-0.444***	0.39
		(-0.73)	(0.00)	(14.33)	(-3.20)	(-2.41)	(-0.44)	(-3.64)	
	Inno	-0.706	1.969	0.529***	-0.151**	-0.822***	-0.410***	0.053	0.29
		(-0.23)	(0.66)	(9.51)	(-1.97)	(-6.51)	(-4.06)	(0.35)	
	Diff	1.235	1.967	-0.122**	0.049	-0.574***	-0.374***	0.496***	0.06
		(0.52)	(0.72)	(-2.41)	(0.71)	(-5.00)	(-4.07)	(3.65)	
Size	Non-Inno	-4.236	-8.669***	0.254***	-0.887***	0.138	0.753***	0.021	0.42
		(-1.61)	(-3.67)	(5.78)	(-14.69)	(1.38)	(9.42)	(0.18)	
	Inno	-7.544**	-9.464***	-0.039	-1.336***	0.027	0.788***	0.141	0.66
		(-2.46)	(-4.48)	(-0.99)	(-24.76)	(0.31)	(11.03)	(1.34)	
	Diff	-3.308*	-0.795	-0.292***	-0.450***	-0.110	0.035	0.120	0.29
		(-1.68)	(-0.41)	(-8.04)	(-9.00)	(-1.34)	(0.54)	(1.23)	
B/M	Non-Inno	4.283**	1.506	-0.024	0.174***	1.029***	-0.587***	0.173**	0.41
		(2.33)	(0.90)	(-0.78)	(4.10)	(14.64)	(-10.41)	(2.08)	
	Inno	5.290**	2.442	0.078*	0.337***	1.142***	-0.596***	0.011	0.33
		(2.28)	(1.09)	(1.87)	(5.91)	(12.12)	(-7.89)	(0.10)	
	Diff	1.006	0.936	0.102**	0.163***	0.113	-0.009	-0.162	0.05
		(0.49)	(0.40)	(2.33)	(2.72)	(1.14)	(-0.11)	(-1.38)	
Invest	Non-Inno	-4.556***	0.440	0.022	-0.194***	-0.847***	0.119***	-0.164***	0.47
		(-3.74)	(0.42)	(1.11)	(-7.27)	(-19.20)	(3.37)	(-3.14)	
	Inno	-2.679	2.524	0.005	-0.129***	-1.588***	0.033	0.192**	0.44
		(-1.30)	(1.39)	(0.14)	(-2.77)	(-20.72)	(0.54)	(2.12)	
	Diff	1.877	2.083	-0.017	0.066	-0.741***	-0.086	0.356***	0.10
		(0.96)	(0.95)	(-0.42)	(1.18)	(-8.04)	(-1.16)	(3.27)	
Profit	Non-Inno	6.225**	-1.737	-0.116***	-0.315***	0.232**	1.211***	0.099	0.47
		(2.35)	(-0.76)	(-2.74)	(-5.38)	(2.40)	(15.63)	(0.86)	
	Inno	-3.650	-7.176**	-0.213***	-0.541***	0.071	0.794***	0.134	0.31
		(-1.20)	(-2.39)	(-3.82)	(-7.06)	(0.56)	(7.82)	(0.89)	
	Diff	-9.876***	-5.439	-0.097	-0.227***	-0.161	-0.417***	0.035	0.03
		(-3.44)	(-1.63)	(-1.56)	(-2.66)	(-1.14)	(-3.70)	-0.21	

The table shows the average excess returns of innovative and non-innovative firms sorted on common firm characteristics as well as the results of regressing the portfolio returns on a constant and HMXZ q5-factors. Stocks are labeled as innovators and non-innovators at the end of June in each year and then sorted into five portfolios within the two groups. Innovative firms are firms that have at least three patents over the last three years and one patent over the last year at the time of portfolio formation. The table shows the returns of a long-short portfolio that goes long the highest quintile and short the lowest quintile. The sample period is 1967-2021. More details can be found in the caption of Table 4.11.

Table 4.12: Pricing characteristics-sorted portfolios in innovative vs. non-innovative firms with the q5-factor model

in the context of technology and growth, which shape the behavior of firms and the development of economies for decades into the future.

Chapter 5

Conclusion

This thesis is a collection of three essays that investigate how financial markets interact with information about future earnings.

The first chapter shows that analyst report texts contain information about future earnings. A simple, dictionary-based tone measure captures information about future earnings that is not reflected in the numerical earnings forecasts in the same reports. The publication of analyst reports causes significant stock price movements on the publication day. However, analyst tone does not predict future stock returns, suggesting that the initial publication reaction adequately incorporates the tonal information into prices. The chapter highlights that analysts use the report text to convey information and that financial market participants pay close attention to the textual information in these documents.

The second chapter develops a new methodology to study pricing-relevant information from analyst report texts. In contrast to the dictionary-based method of the first chapter, which studies a particular measure of linguistic sentiment and the embedded information about future earnings, the deep learning method presented in this chapter explicitly extracts pricing-relevant information from the reports. I demonstrate that not all information in analyst reports is priced correctly upon publication. The deep neural network is able to predict earnings announcement returns days or weeks after the publication of the reports. Initial price reactions and drifts - or the absence thereof - suggest that markets initially react insufficiently to particular information in the reports and that this insufficient initial reaction is only corrected on the date of the actual earnings announcement following the report publication. The identified predictors of future returns can be used to predict returns several years out-of-sample. Without retraining, the neural network eventually loses its predictive power, suggesting that markets slowly learn from past pricing mistakes.

The third chapter shows that publicly available information about firms' innovation activities is an important determinant of the cross-section of stock returns. Firms with high market patenting intensity, defined as the ratio of the number of patents assigned to the firm and its market capitalization, earn a significant return premium several years after the patent grant date. Not only do common factor pricing models fail to explain this return premium, but some of the underlying cross-sectional relationships that these factor pricing models are based on do not exist among innovative firms. In particular, innovative firms with low profitability still earn high returns in contrast to the well-known profitability anomaly. Moreover, there is no significant relationship between tangible investments and returns in innovative firms. Portfolios sorted on patenting intensity are strikingly stable, producing abnormal five-factor returns for a full decade after portfolio formation.

Bibliography

Abarbanell, Jeffery S., and Victor L. Bernard, 1992, Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behavior, *Journal of Finance* 47, 1181–1207.

Aghion, Philippe, and Peter Howitt, 1992, A model of growth through creative destruction, *Econometrica* 60, 323–351.

Antweiler, Werner, and Murray Z. Frank, 2004, Is all that talk just noise? The information content of internet stock message boards, *Journal of Finance* 59, 1259–1294.

Azimi, Mehran, and Anup Agrawal, 2021, Is positive sentiment in corporate annual reports informative? Evidence from deep learning, *Review of Asset Pricing Studies* 11, 762–805.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller, 2010, How to explain individual classification decisions, *Journal of Machine Learning Research* 11, 1803–1831.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, 2014, Neural machine translation by jointly learning to align and translate, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, arXiv:1409.0473.

Bali, Turan G., Amit Goyal, Dashan Huang, Fuwei Jiang, and Quan Wen, 2020, The cross-sectional pricing of corporate bonds using big data and machine learning, Swiss Finance Institute Research Paper 20-110, Swiss Finance Institute.

Ball, Ray, and Philip Brown, 1968, An Empirical Evaluation of Accounting Income Numbers, *Journal of Accounting Research* 6, 159–178.

Beaver, William H., 1968, The information content of annual earnings announcements, *Journal of Accounting Research* 6, 67–92.

Bena, Jan, and Lorenzo Garlappi, 2020, Corporate innovation and returns, *Review of Corporate Finance Studies* 9, 340–383.

Berk, Jonathan B, 1995, A critique of size-related anomalies, *Review of Financial Studies* 8, 275–286.

Bernard, Victor L., and Jacob K. Thomas, 1989, Post-earnings-announcement drift: Delayed price response or risk premium?, *Journal of Accounting Research* 27, 1–36.

Bernard, Victor L., and Jacob K. Thomas, 1990, Evidence that stock prices do not fully reflect the implications of current earnings for future earnings, *Journal of Accounting and Economics* 13, 305–340.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni, 2021, Bond risk premiums with machine learning, *Review of Financial Studies* 34, 1046–1089.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* 5, 135–146.

Bondt, Werner F. M. De, and Richard H. Thaler, 1990, Do Security Analysts Overreact?, *American Economic Review* 80, 52–57.

Bouchaud, Jean-Philippe, Philipp Krüger, Augustin Landier, and David Thesmar, 2019, Sticky expectations and the profitability anomaly, *Journal of Finance* 74, 639–674.

Bound, John, Clint Cummins, Zvi Griliches, Bronwyn H. Hall, and Adam B. Jaffe, 1982, Who does R&D and who patents?, Technical report, National Bureau of Economic Research.

Bradshaw, Mark T., 2011, Analysts' Forecasts: What Do We Know after Decades of Work?, SSRN Scholarly Paper ID 1880339, Social Science Research Network, Rochester, NY.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2019, Bayesian solutions for the factor zoo: We just ran two quadrillion models, SSRN Scholarly Paper ID 3481736, Social Science Research Network, Rochester, NY.

Chan, Louis K.C., Josef Lakonishok, and Theodore Sougiannis, 2001, The stock market valuation of research and development expenditures, *Journal of Finance* 56, 2431–2456.

Chang, Tom Y., Samuel M. Hartzmark, David H. Solomon, and Eugene F. Soltes, 2017, Being surprised by the unsurprising: Earnings seasonality and stock returns, *Review of Financial Studies* 30, 281–323.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2020, Deep learning in asset pricing, SSRN Scholarly Paper ID 3350138, Social Science Research Network.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio, 2014a, On the properties of neural machine translation: Encoder-decoder approaches, in *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, arXiv:1409.1259.

Cho, Kyunghyun, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, 2014b, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, arXiv:1406.1078.

Cohen, Lauren, Karl Diether, and Christopher Malloy, 2013, Misvaluing innovation, *Review of Financial Studies* 26, 635–666.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *Journal of Finance* 75, 1371–1415.

Daniel, Kent, David Hirshleifer, and Lin Sun, 2020, Short- and long-horizon behavioral factors, *Review of Financial Studies* 33, 1673–1736.

De Franco, Gus, Ole-Kristian Hope, Dushyantkumar Vyas, and Yibin Zhou, 2015, Analyst Report Readability, *Contemporary Accounting Research* 32, 76–104.

Dellavigna, Stefano, and Joshua M. Pollet, 2009, Investor Inattention and Friday Earnings Announcements, *Journal of Finance* 64, 709–749.

Eberhart, Allan C., William F Maxwell, and Akhtar R. Siddique, 2004, An examination of long-term abnormal stock returns and operating performance following R&D increases, *Journal of Finance* 59, 623–650.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al., 2004, Least angle regression, *The Annals of Statistics* 32, 407–499.

Elliott, John, Gordon Richardson, Thomas Dyckman, and Roland Dukes, 1984, The impact of SFAS No. 2 on firm expenditures on research and development: Replications and extensions, *Journal of Accounting Research* 22, 85–102.

Engelberg, Joseph, R. David McLean, and Jeffrey Pontiff, 2018, Anomalies and news, *Journal of Finance* 73, 1971–2001.

Engelberg, Joseph, R. David McLean, and Jeffrey Pontiff, 2019, Analysts and anomalies, *Journal of Accounting and Economics* 101249.

Fama, Eugene F., 1965, The behavior of stock-market prices, *Journal of Business* 38, 34–105.

Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 1995, Size and book-to-market factors in earnings and returns, *Journal of Finance* 50, 131–155.

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

Fama, Eugene F., and Kenneth R. French, 2018, Choosing factors, *Journal of Financial Economics* 128, 234–252.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020a, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.

Feng, Guanhao, Nick Polson, and Jianeng Xu, 2020b, Deep learning in characteristics-sorted factor models, SSRN Scholarly Paper ID 3243683, Social Science Research Network.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

Gompers, Paul, Anna Kovner, Josh Lerner, and David Scharfstein, 2008, Venture capital investment cycles: The impact of public markets, *Journal of Financial Economics* 87, 1–23.

Grossman, Sanford J, and Joseph E Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American Economic Review* 70, 393–408.

Gu, Feng, 2005, Innovation, future earnings, and market efficiency, *Journal of Accounting, Auditing & Finance* 20, 385–418.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

Gultekin, Mustafa, Abena Owusu, Thomas Shohfi, and Majeed Simaan, 2019, How have stock markets responded to 35 years of analyst reports? Evidence from machine learning and textual analysis, Working Paper.

Hall, Bronwyn H., 1993, The stock market's valuation of R&D investment during the 1980's, *American Economic Review* 83, 259–264.

Hall, Bronwyn H., and Robert E. Hall, 1993, The value and performance of US corporations, *Brookings Papers on Economic Activity* 1993, 1–49.

Heaton, J. B., N. G. Polson, and J. H. Witte, 2017, Deep learning for finance: deep portfolios, *Applied Stochastic Models in Business and Industry* 33, 3–12.

Hillert, Alexander, Heiko Jacobs, and Sebastian Müller, 2014, Media makes momentum, *Review of Financial Studies* 27, 3467–3501.

Hinton, Geoffrey E, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, 2012, Improving neural networks by preventing co-adaptation of feature detectors, *CoRR* arXiv:1207.0580.

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li, 2013, Innovative efficiency and stock returns, *Journal of Financial Economics* 107, 632–654.

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li, 2018, Innovative originality, profitability, and stock returns, *Review of Financial Studies* 31, 2553–2605.

Hirshleifer, David, Sonya Seongyeon Lim, and Siew Hong Teoh, 2009, Driven to distraction: Extraneous events and underreaction to earnings news, *Journal of Finance* 64, 2289–2325.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White, 1989, Multilayer feedforward networks are universal approximators, *Neural networks* 2, 359–366.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2021, An augmented q-factor model with expected growth, *Review of Finance* 25, 1–41.

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

Hou, Kewei, Chen Xue, and Lu Zhang, 2020a, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020b, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Hsieh, Chia-Chun, Kai Wai Hui, and Yao Zhang, 2016, Analyst Report Readability and Stock Returns, *Journal of Business Finance & Accounting* 43, 98–130.

Huang, Allen H., Reuven Lehavy, Amy Y. Zang, and Rong Zheng, 2017, Analyst information discovery and interpretation roles: A topic modeling approach, *Management Science* 64, 2833–2855.

Huang, Allen H., Amy Y. Zang, and Rong Zheng, 2014, Evidence on the information content of text in analyst reports, *The Accounting Review* 89, 2151–2180.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou, 2019, Manager sentiment and stock returns, *Journal of Financial Economics* 132, 126–149.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, 2017, Bag of tricks for efficient text classification, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, 427–431.

Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu, 2020, Predicting returns with text data, NBER Working Paper 27950, National Bureau of Economic Research.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, 2021, Measuring technological innovation over the long run, Technical Report 3.

Kim, Yoon, 2014, Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, arXiv:1408.5882.

Kingma, Diederik P, and Jimmy Ba, 2014, Adam: A method for stochastic optimization, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, arXiv:1412.6980.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, 2017, Self-normalizing neural networks, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 972–981, arXiv:1706.02515.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, 2017, Technological innovation, resource allocation, and growth, *Quarterly Journal of Economics* 132, 665–712.

Kuhnen, Camelia M., 2015, Asymmetric Learning from Financial Information, *Journal of Finance* 70, 2029–2062.

Lev, Baruch, and Theodore Sougiannis, 1996, The capitalization, amortization, and value-relevance of R&D, *Journal of Accounting and Economics* 21, 107–138.

Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar, 2018, Hyperband: A novel bandit-based approach to hyperparameter optimization, *Journal of Machine Learning Research* 18, 1–52.

Liberti, José María, and Mitchell A. Petersen, 2019, Information: Hard and Soft, *Review of Corporate Finance Studies* 8, 1–41.

Lochstoer, Lars A., and Paul C. Tetlock, 2020, What drives anomaly returns?, *Journal of Finance* 75, 1417–1455.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill Mcdonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning, 2015, Effective approaches to attention-based neural machine translation, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, arXiv:1508.04025.

Martineau, Charles, 2019, The evolution of financial market efficiency: Evidence from earnings announcements, SSRN Scholarly Paper ID 3111607, Social Science Research Network, Rochester, NY.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013a, Efficient estimation of word representations in vector space, in *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, arXiv:1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 2013b, Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, arXiv:1310.4546.

Nagel, Stefan, 2005, Short sales, institutional investors and the cross-section of stock returns, *Journal of Financial Economics* 78, 277–309.

Novy-Marx, Robert, 2013a, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

Novy-Marx, Robert, 2013b, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio, 2013, On the difficulty of training recurrent neural networks, in *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, 1310–1318, arXiv:1211.5063.

Pástor, L'uboš, and Pietro Veronesi, 2006, Was there a nasdaq bubble in the late 1990s?, *Journal of Financial Economics* 81, 61–100.

Pástor, L'uboš, and Pietro Veronesi, 2009, Technological revolutions and stock prices, *American Economic Review* 99, 1451–83.

Peters, Ryan H., and Lucian A. Taylor, 2017, Intangible capital and the investment-q relation, *Journal of Financial Economics* 123, 251–272.

Price, S. McKay, James S. Doran, David R. Peterson, and Barbara A. Bliss, 2012, Earnings conference calls and stock returns: The incremental informativeness of textual tone, *Journal of Banking & Finance* 36, 992–1011.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, 2016, "Why should I trust you?" Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Romer, Paul, 1986, Increasing returns and long-run growth, *Journal of Political Economy* 94, 1002–1007.

Romer, Paul, 1990, Endogenous technological change, *Journal of Political Economy* 98, S71–S102.

Savor, Pavel, and Mungo Wilson, 2016, Earnings announcements and systematic risk, *Journal of Finance* 71, 83–138.

Schmeling, Maik, and Christian Wagner, 2019, Does central bank tone move asset prices?, *Available at SSRN 2629978* .

Schumpeter, Joseph, 1911, *The Theory of Economic Development* (Harvard University Press).

Schuster, M., and K. K. Paliwal, 1997, Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing* 45, 2673–2681.

Schwert, G. William, 2002, Stock volatility in the new millennium: how wacky is Nasdaq?, *Journal of Monetary Economics* 49, 3–26.

Shumway, Tyler, 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.

Sims, Christopher A, 2003, Implications of rational inattention, *Journal of Monetary Economics* 50, 665–690.

Solow, Robert, 1957, Technical change and the aggregate production function, *Review of Economics and Statistics* 39, 312–320.

Srinivas, Niranjan, Andreas Krause, Sham M. Kakade, and Matthias Seeger, 2012, Gaussian process optimization in the bandit setting: No regret and experimental design, *IEEE Transactions on Information Theory* 58, 3250–3265, arXiv:0912.3995.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *Journal of Finance* 63, 1437–1467.

Titman, Sheridan, K.C. John Wei, and Feixue Xie, 2004, Capital investments and stock returns, *Journal of Financial and Quantitative Analysis* 39, 677–700.

Tobin, James, 1958, Estimation of relationships for limited dependent variables, *Econometrica* 24–36.

Tversky, Amos, and Daniel Kahneman, 1973, Availability: A heuristic for judging frequency and probability, *Cognitive Psychology* 5, 207–232.

Tversky, Amos, and Daniel Kahneman, 1974, Judgment under Uncertainty: Heuristics and Biases, *Science* 185, 1124–1131.

van Binsbergen, Jules H., Martijn Boons, Christian C. Opp, and Andrea Tamoni, 2021, Dynamic asset (mis)pricing: Build-up vs. resolution anomalies, Technical report, SSRN.

van Binsbergen, Jules H., and Christian C. Opp, 2019, Real anomalies, *Journal of Finance* 74, 1659–1706.

van Binsbergen, Jules H. van, Xiao Han, and Alejandro Lopez-Lira, 2020, Man vs. machine learning: The term structure of earnings expectations and conditional biases, NBER Working Paper 27843, National Bureau of Economic Research.

Welch, Ivo, and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* 21, 1455–1508.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, 2016, Hierarchical attention networks for document classification, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Zou, Hui, and Trevor Hastie, 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320.

Appendix A

Appendix to Chapter 2

A.1 Matching

A.1.1 Target price based

1. Infer target prices from reports
   - remove currency symbols from corpus ("$", "US$", "USD")
   - remove text in parentheses (to remove annotations, such as target price dates)
   - on the first page of the report, search for the key words "target price", "price target", or "target" followed by a number, potentially preceded by a colon. This number is the target price estimate

2. Match reports with IBES target price file on firm-date-target price to obtain *contributor-estimid* link

3. Verify *contributor-estimid* links
   - For non-masked *estimid*s (all but PRMDN*), check whether *estimid* is roughly similar to the contributor name, or known former names of this contributor
   - The sum of all verified contributors should cover a large proportion (e.g. >90%) of linked reports for each *estimid*

4. Merge with *estimid-emaskcd* link from IBES recommendations file

Note that *estimid-contributor* links are not unique due to inconsistent spelling in Thomson One as well as mergers and acquisitions. For example, *estimid* "BEAR" should be linked to the *contributor*s "BEAR, STEARNS & CO., INC.", "BEAR, STEARNS MORNING MEETING NOTES", and "BEAR STEARNS AND CO INC". Similarly, the *contributor* "CIBC CAPITAL MARKETS CORP." should be linked to the *estimid*s "OPPEN" and "WOODGUND". We keep all possible links at this point and take care of ambiguous links later on.

A.1.2 Hand matching

For each *contributor*:

1. Randomly select report from this *contributor*
    - Manually read earnings forecasts for the next two fiscal years from the report (use fiscal quarters if annual estimates are not available)
    - Look for matching entry in IBES Details (i.e. a emaskcd with matching firm-date-estimate for both forecasts)
    - Record emaskcd of matched entry (entries)

2. Repeat the previous step up to ten times until the same *emaskcd* has been matched at least five times

3. The *emaskcd-contributor* match is considered invalid if less than five matches were found, or if multiple *emaskcd*s have been matched more than twice with this contributor[1]

A.2 Text cleaning

The first step of the text cleaning procedure seeks to remove all text blocks that do not contain structured text or paragraphs that do not contain analyst opinions.

1. Remove report pages do not contain any elements other than the following, according to the table of contents: *'Disclosure', 'Disclaimer', 'Analyst Certification', 'Report Key', 'Page is Blank/No Information','Company Description', 'Company Profile', 'Table/Chart', 'Graph', 'Table'*. Note that the title in the table of contents can be followed by the expression *'(CONT)'* or a set of Roman or Arabic numerals.

2. Split reports into paragraphs

3. Remove non-ascii characters

4. Remove trailing special characters

5. Remove repeated whitespaces characters

[1]Matching multiple *emaskcd*s is rare, thus most *emaskcd-contributor* links are based on five out of five matches.

6. Remove trailing whitespace characters

7. Remove paragraphs that match any of the following expressions ignoring capitalization. The caret symbol indicates that the expression must be found at the beginning of the paragraph: *'does and seeks to do business with companies covered in', 'has been an investment banking client of', 'has received compensation for investment banking services', '^analyst certification', '^disclosure', 'disclosures and analyst certifications', '^important disclosure', '^company description', '^this report was produced by', '^this report has been prepared by', '^disclaimer', '^global disclaimer', '^customers of the firm in the United States can receive independent, third-party research', 'all rights reserved'*

The second step is to extract relevant words from the report text.

1. Convert all characters to lowercase

2. Remove possessive *'s*

3. Replace special characters with white spaces

4. Lemmatize words using the *spaCy* lemmatizer

A.3 Robustness checks

A.3.1 Predictability of forecast errors

Level vs change in analyst tone

In Table A.1, we investigate whether the level or the change in analyst tone is responsible for the predictability of forecast errors. Univariate regressions suggest that while both the level and the change in analyst tone have a positive correlation with forecast errors, the level has much higher explanatory power than the change. Controlling for the level of post-announcement analyst tone S_t, the change in analyst tone around the announcement date predicts $t+1Q$ and $t+4Q$ forecast errors with a negative sign. However, the relationship is statistically weak for $t+1Q$ errors, and economically small for all forecast periods relative to the level effect. This is in line with Huang et al. (2014), who find that the analyst opinions rather than changes in their opinions explain the market reaction on analyst report announcement dates.

	FE_t^{t+1Q}	FE_t^{t+1Q}	FE_t^{t+1Q}	FE_t^{t+4Q}	FE_t^{t+4Q}	FE_t^{t+4Q}
$Tone_t$	0.074***		0.091***	0.089***		0.117***
	(15.71)		(12.59)	(14.90)		(12.98)
$Tone_t - \text{Tone}_{t^-}$		0.029***	-0.009*		0.021***	-0.028***
		(6.74)	(-1.70)		(4.25)	(-5.02)
Fixed Effects	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)
N	86,821	49,886	49,886	73,225	43,259	43,259
R^2	0.117	0.134	0.138	0.145	0.172	0.179
Within R^2	0.005	0.001	0.006	0.007	0.000	0.008

This table reports regressions of post-announcement forecast errors onto post-announcement analyst tone as well as the change in analyst tone. All variables are measured at the firm level, therefore, firm subscripts are omitted for ease of notation. Post-announcement forecast errors $E_t^{t+\tau}$ are realized actuals for period $t+\tau$ (shown in the superscript) minus consensus expectations measured after the earnings announcement in period t scaled by the stock price 46 days prior to the announcement. Post-announcement consensus expectations are the average of individual analyst expectations measured within 45 days after the announcement. Post-announcement analyst tone $Tone_t$ is the average tone of the reports published along with the selected individual expectations. Similarly, pre-announcement analyst tone $Tone_{t^-}$ is the average tone of the reports published along with pre-announcement expectations within 45 days prior to the announcement. $Tone_t - Tone_{t^-}$ is the difference between post and pre-announcement analyst tone. See Section 2.2.3 for details on the measurement and aggregation of analyst tone. Regressions include firm (f) and year-quarter (t) fixed effects. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis.

Table A.1: Sentiment vs sentiment changes and future errors (post-announcement)

A.3.2 Portfolio sorting

	low	med	high	low-high
Intercept	0.331***	0.157***	-0.032	0.331***
	(4.56)	(2.60)	(-0.58)	(3.73)
Mkt-RF	1.090***	1.020***	1.036***	0.039
	(36.96)	(37.97)	(42.79)	(1.02)
SMB	0.476***	0.289***	0.394***	0.128*
	(8.61)	(6.50)	(9.31)	(1.86)
HML	0.450***	0.251***	0.094**	0.339***
	(8.72)	(5.80)	(2.35)	(5.31)
N	591	585	662	544
$Adj.R^2$	0.75	0.75	0.76	0.05

This table shows the returns of a tone-based trading strategy identical to the post-announcement strategy in Table 2.7, benchmarked against a Fama-French (1993) three-factor model, omitting the Momentum factor.

Table A.2: FF3 risk-adjusted returns of an analyst tone-based trading strategy

A.3.3 Predicting forecast errors

	C_t/P_{t^-} (%)	C_t/P_{t^-} (%)	C_t/P_{t^-} (%)	C_t/P_{t^-} (%)
$\mathbb{E}_{t^-}[C_t]/P_{t^-}$ (%)	0.959***	0.913***	0.960***	0.914***
	(269.46)	(190.91)	(265.31)	(189.05)
$Tone_{t^-}$			0.005***	0.009***
			(3.59)	(7.70)
Constant	0.095***		0.094***	
	(16.52)		(15.77)	
Fixed Effects		(f,t)		(f,t)
N	78,482	78,275	78,482	78,275
R^2	0.853	0.876	0.853	0.876
Within R^2		0.751		0.752

Panel A: Current earnings

	C_{t+4Q}/P_{t^-} (%)	C_{t+4Q}/P_{t^-} (%)	C_{t+4Q}/P_{t^-} (%)	C_{t+4Q}/P_{t^-} (%)
$\mathbb{E}_{t^-}[C_{t+4Q}]/P_{t^-}$ (%)	0.752***	0.624***	0.758***	0.629***
	(68.32)	(48.50)	(69.46)	(48.25)
$Tone_{t^-}$			0.025***	0.027***
			(4.98)	(7.78)
Constant	0.250***		0.240***	
	(19.05)		(17.73)	
Fixed Effects		(f,t)		(f,t)
N	46,466	46,188	46,466	46,188
R^2	0.492	0.617	0.494	0.618
Within R^2		0.283		0.285

Panel B: Four quarters ahead earnings

This table report regression results of future earnings on earnings expectations. In Panel A, the dependent variable C_t/P_{t^-} is the period-t earnings scaled by price. The independent variables include the pre-announcement earnings expectations $\mathbb{E}[C_t]/P_{t^-}$ and the pre-announcement analyst tone $Tone_{t^-}$ as defined in Section 2.2.3. In Panel B, the dependent variable C_{t+4Q}/P_{t^-} is the four-quarter ahead earnings scaled by price, and the independent variable $\mathbb{E}[C_{t+4Q}]/P_{t^-}$ is the expectation of the four-quarter ahead earnings prior to the period-t earnings announcement. Regressions include firm (f) and year-quarter (t) fixed effects as denoted in the fixed effects row. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parentheses.

Table A.3: Error components

	$FE_{t^-}^{t}$	$FE_{t^-}^{t}$	$FE_{t^-}^{t}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$	$FE_{t^-}^{t+4Q}$
$Tone_{t^-}$	0.033***		0.033***	0.090***		0.099***	0.095***
	(5.26)		(5.23)	(8.73)		(10.12)	(10.22)
Rec_{t^-}		0.013	-0.003		-0.020	-0.066***	-0.066***
		(1.00)	(-0.19)		(-0.93)	(-3.03)	(-3.20)
$FE_{t^-}^{t}$							0.171***
							(24.13)
Constant	0.006	-0.041	0.016	-0.020	0.070	0.226**	0.225***
	(0.44)	(-0.84)	(0.31)	(-0.94)	(0.78)	(2.54)	(2.66)
N	66,773	66,773	66,773	34,681	34,681	34,681	34,681
R^2	0.001	0.000	0.001	0.006	0.000	0.007	0.038

Same as Table 2.2 but without fixed effects.

Table A.4: Pre-announcement sentiment and forecast errors, no fixed effects

	$FE_{t^-}^{t}$	$FE_{t^-}^{t+1Q}$	$FE_{t^-}^{t+1Q}$	$FE_{t^-}^{t+2Q}$	$FE_{t^-}^{t+3Q}$	$FE_{t^-}^{t+4Q}$
$Tone_{t^-}$	0.056***	0.098***	0.093***	0.103***	0.099***	0.097***
	(10.23)	(16.20)	(15.22)	(14.22)	(12.11)	(11.51)
$FE_{t^-}^{t}$			0.192***	0.132***	0.103***	0.150***
			(26.05)	(18.72)	(13.94)	(22.00)
Fixed Effects	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)	(f,t)
N	68,241	65,777	51,128	46,098	40,630	34,389
R^2	0.134	0.113	0.164	0.162	0.176	0.208
Within R^2	0.002	0.006	0.045	0.027	0.019	0.031

Same as Table 2.2 but with additional horizons

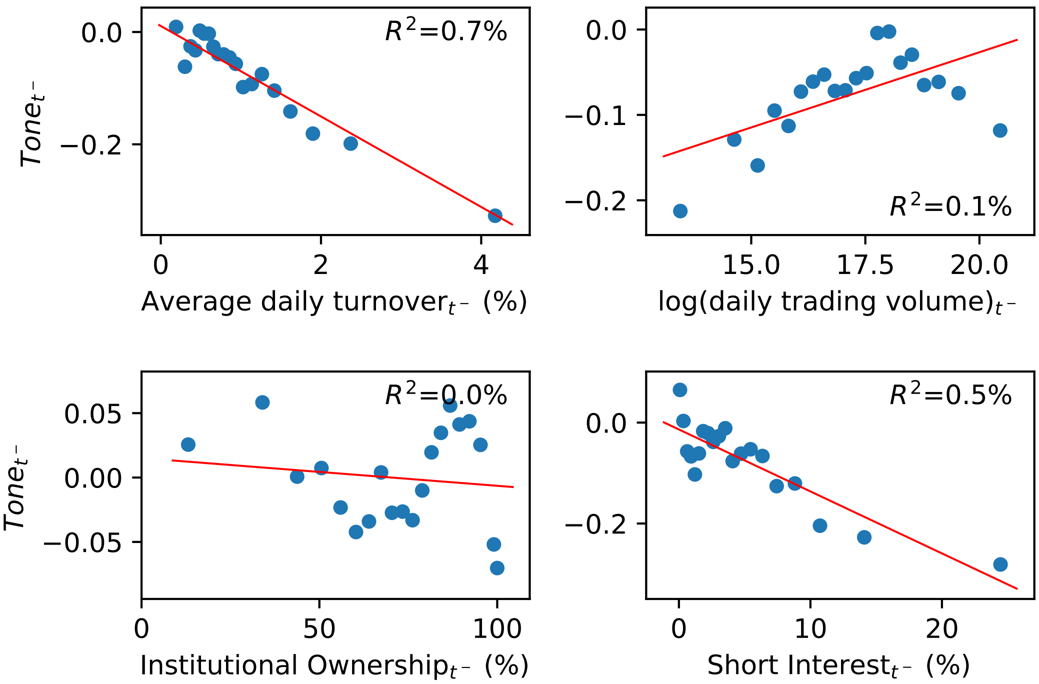
Table A.5: Pre-announcement sentiment and forecast errors, additional horizons

	Full Sample	Excl. recessions	Recessions
	$FE^{t}_{t^-}$	$FE^{t}_{t^-}$	$FE^{t}_{t^-}$
$Tone_{t^-}$	0.056***	0.055***	0.053**
	(10.23)	(9.65)	(2.42)
N	68,241	62,306	5,557
R^2	0.134	0.140	0.338

Same as Table 2.2 but splitting the sample into recession and non-recession periods. Recession periods are defined by the NBER recession indicator and include Jul 1990 to Mar 1991, Mar 2001 to Nov 2001, and Dec 2007 to Jun 2009.

Table A.6: Pre-announcement sentiment and forecast errors, business cycle subsamples

A.3.4 Miscellaneous



The figure shows binscatter plots of pre-announcement analyst tone $Tone_{t^-}$ as defined in Eq. (2.4) against various control variables. For each panel, the variable on the horizontal axis is sorted into 20 equally-sized bins. Each dot represents the mean analyst tone and the mean of the binned variable for one bin. The red line represents a linear fit of the underlying data. The R^2 of this linear regression is shown in each panel. The top-left panel plots analyst tone against average daily turnover in the pre-announcement window [t-45,t-1]. Daily turnover is the fraction of outstanding shares that was been traded on a single day. Both the number of shares traded and the number of shares outstanding are from CRSP. The top-right panel plots analyst tone against the logarithm of the average daily dollar trading volume, approximated by the number of shares traded times the daily closing price from CRSP. The bottom-left panel plots analyst tone against the percentage of shares owned by institutional shareholders. Institutional ownership is measured by the latest institutional ownership data available in the Thomson Reuters 13f database within [t-100,t-1]. The bottom-right panel plots analyst tone against the fraction of shares shorted. The number of shares shorted is from Compustat, while the number of shares outstanding is from CRPS. For each firm and date, we use the latest short interest figure available within [t-45,t-1].

Figure A.1: Analyst tone and additional control variables

	$Tone_{t^-}$							
$Tone_{(t-1)^-}$	0.561***							
	(79.67)							
$\frac{\mathbb{E}_{t-1}[C^t]-\mathbb{E}_{t^-}[C^t]}{P_{t-1}}$		0.142***						
		(18.40)						
Rec_{t^-}			0.532***					
			(26.97)					
TPIR_{t^-}				-0.095				
				(-1.37)				
$\log(\text{market cap})_{t^-}$					0.017**			
					(2.52)			
BM_{t^-}						-0.727***		
						(-29.26)		
Volatility_{t^-}							-0.703***	
							(-6.32)	
Return_{t^-}								0.899***
								(3.12)
N	106347	61389	116163	75070	81378	80958	62288	62288
R^2	0.338	0.027	0.078	0.000	0.001	0.089	0.030	0.023

Regressions as shown in Fig. 2.4. Standard errors are clustered by firm and year-quarter. T-statistics are shown in parenthesis. *,**, and *** indicate p-values of less than 10%, 5%, and 1%, respectively.

Table A.7: Linear relationship between pre-announcement analyst tone on other pre-announcement variables.

Appendix B

Appendix to Chapter 3

B.1 Neural network details

B.1.1 Embeddings

Words are embedded into a low-dimensional vector space using the *fasttext* algorithm of Bojanowski et al. (2017) based on the skipgram algorithm with negative sampling as in Mikolov et al. (2013a). The objective of the skipgram algorithm is to predict the neighboring words of each word in the text corpus from a low-dimensional vector representation. This vector representation, called the word embedding, is learned during the training process. I create two randomly initialized matrices of word representation, $W_e \in \mathbf{R}^{E \times V}$ and $U \in \mathbf{R}^{V \times E}$. The first matrix W_e maps the one-hot encoded word vectors of length V (the size of the vocabulary) to the low-dimensional embedding vector of length E. I use w_i to refer to columns in W_e and u_i to refer to rows in U. These vectors are often referred to as input and output vectors. Only the input vectors will be used in the downstream task. A simple way to generate predictions of context words would be to maximize the likelihood of the observed context words, i.e. for M reports of length N_m we would maximize the likelihood

$$\prod_{m=1}^{M} \prod_{i=1}^{N_m} \prod_{c \in C_i} p(c|i)$$

where C_i is the set of context words of word i, and $p(c|i)$ could be parametrized with W_e and V and a softmax function. However, while mathematically simple, this specification is computationally expensive, since it requires us to calculate gradients of every possible context word (i.e. every word in the vocabulary) at every step. Instead, I use negative sampling to drastically reduce the number of candidate context that we evaluate for each word i. I do so by approximating the likelihood maximization by a series of binary classification tasks. For each focus word i and observed context words C_i, I randomly sample five words $\tilde{C}_i$ from the vocabulary that do not appear in C_i. Formally, we can write the objective function as

$$\prod_{m=1}^{M}\prod_{i=1}^{N_m}\prod_{c\in C_i} p(c|i) \prod_{\tilde{c}\in\tilde{C}_{i,c}} (1-p(\tilde{c}|i))$$

I model the probability $p(c|i)$ with the logistic function,

$$p(c|i) = \frac{1}{1+\exp(-s(c,i))}$$

where $s(c,i)$ is a score function that converts the input and output vector representations corresponding to c and i to a scalar measure of vector similarity via the dot product,

$$s(c,i) = w_i^\top u_c$$

Note that we use a different vector representation for a given word when it appears as input (w) than when it appears as a context word (u). So far, we ignored the internal structure of words. If we were to treat w as a free parameter as in the basic skipgram model of Mikolov et al. (2013b), we would neglect any type of morphological information that can be used to relate words to each other. Bojanowski et al. (2017) suggest to further parametrize the input word vectors w by modeling them as linear combination of subword vectors. In particular, I represent each word as the original word plus a bag of character n-grams. Following the original literature, I set n to be between 3 and 6, and add the boundary symbols $<$ and $>$ to the beginning and the end of the word to distinguish prefixes and suffixes from other n-grams. For example, the word *finance* will be represented as a collection of the n-grams

{ <fi, fin, ina, nan, anc, nce, ce>,
<fin, fina, inan, nanc, ance, nce>,
<fina, finan, inanc, nance, ance>
<finan, financ, inance, nance>,
<finance>}

I refer to the collection of n-grams for word i as G_i. I then associate a vector representation $z_g \in \mathbf{R}^E$ to each unique n-gram g and obtain the word representation w_i by taking the sum of all the subword representations,

$$w_i = \sum_{g\in G_i} z_g.$$

Finally, I can define the objective function as

$$\begin{aligned}
&\arg\max_{\mathscr{X},U} \prod_{m=1}^{M}\prod_{i=1}^{N_m}\prod_{c\in C_i} p(c|i) \prod_{\tilde{c}\in\tilde{C}_{i,c}} (1-p(\tilde{c}|i)) \\
=&\arg\max_{\mathscr{X},U} \sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{c\in C_i} \left(\log p(c|i) + \sum_{\tilde{c}\in\tilde{C}_{i,c}} \log(1-p(\tilde{c}|i))\right) \\
=&\arg\max_{\mathscr{X},U} \sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{c\in C_i} \left(\log \frac{1}{1+\mathsf{exp}(-w_i^\top u_c)} + \sum_{\tilde{c}\in\tilde{C}_{i,c}} \log(1-\frac{1}{1+\mathsf{exp}(-w_i^\top u_{\tilde{c}})}\right) \\
=&\arg\max_{\mathscr{X},U} -\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{c\in C_i} \left(\log(1+\mathsf{exp}(-w_i^\top u_c)) + \sum_{\tilde{c}\in\tilde{C}_{i,c}} \log(1+\mathsf{exp}(w_i^\top u_{\tilde{c}}))\right)
\end{aligned}$$

where $\mathscr{X} \in \mathbf{R}^{E\times|\mathscr{G}|}$ is the matrix of n-gram embeddings (each column is equal to a unique z_g). $\mathscr{G} = \bigcup_{m=1}^{M}\bigcup_{i=1}^{N_m} G_i$ is the collection of unique n-grams over the entire sample. I solve the optimization problem with stochastic gradient descent.

B.1.2 GRU details

The implementation of the GRU follows Cho et al. (2014a). The unit comprises a *hidden state* h_t, which is the output of the unit, and two *gates* z_t and r_t that control the flow of information across time steps. All states and gates are vectors of length H.

For an input sequence $\{w_1,\dots,w_t\}$, the hidden state at step t is a weighted combination of the previous hidden states and a new candidate hidden state. In particular,

$$h_t = z_t \odot \tilde{h}_t + (1-z_t)\odot h_{t-1}$$

where $\tilde{h}_t$ is the candidate hidden state state,

$$\tilde{h}_t = \mathsf{tanh}\left(W_h w_t + U_h(r_t \odot h_{t-1}) + b_h\right)$$

and

$$\begin{aligned}
z_t &= \mathsf{sigm}\left(W_z w_t + U_z h_{t-1} + b_z\right)\\
r_t &= \mathsf{sigm}\left(W_r w_t + U_r h_{t-1} + b_r\right)
\end{aligned}$$

are the gates. $\odot$ is the Hadamart product (element-wise product) operator, sigm is the element-wise sigmoid (logistic) function and tanh is the element-wise hyperbolic tangent function. z_t is the update gate that regulates how much of the previous hidden state gets carried over to the new hidden state. r_t is the reset gate that controls which parts of the previous hidden state enter the new candidate hidden state. h_0 is a vector of zeros.

In the main paper, I use the functional notation $\mathrm{GRU}(\cdot)$ which summarizes the sequence of algebraic operations above. In particular, $\mathrm{GRU}(\cdot)$ is a vector-valued function that outputs h_t, i.e.

$$\mathrm{GRU}(w_1,\ldots,w_t) = \mathrm{GRU}_{\{W_h,U_h,b_h,W_z,U_z,b_z,W_r,U_r,b_r\}}(w_1,\ldots,w_t) = h_t$$

B.1.3 Hyperparameter search space

I restrict the parameter space that is explored by the GP-UCB algorithm to a coarse grid of values that are commonly used in the machine learning literature. The search space is shown in Table B.1.

Parameter	Symbol	Range
Layer dimensions		
Word embedding size	E	[64,128]
GRU hidden units (combined directions)	$2H$	[16, 32, 64]
Report encoder intermediate layer	D_1	[16, 32, ..., 256]
Report encoder output dimension	D_2	[16, 32, ..., 256]
Regularization parameters		
Report encoder word attention W_w l_2-regularizer	λ_{W_w}	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Report encoder word attention u_w l_2-regularizer	λ_{u_w}	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Report encoder W_{h1} l_2-regularizer	$\lambda_{W_{h_1}}$	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Report encoder W_{h2} l_2-regularizer	$\lambda_{W_{h_2}}$	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Word attention output dropout		$[0, 0.1, \ldots, 0.5]$
Report attention W_d l_2-regularizer	λ_{W_d}	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Report attention u_d l_2-regularizer	λ_{u_d}	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$
Report attention output dropout		$[0, 0.1, \ldots, 0.5]$
Final prediction W_v l_2-regularizer	λ_{W_v}	$[0, 10^{-10}, 10^{-9}, \ldots, 10^{-3}]$

Table B.1: Neural network hyperparameter search space

B.2 Benchmark models

B.2.1 Sentiment dictionary

Instead of learning relevant features from the data, the sentiment dictionary approach measures the sentiment in each report by using a sentiment dictionary. The sentiment score is then turned into a return prediction with a simple univariate regression. Sentiment dictionary approaches are popular in the analyst literature. Huang et al. (2014) shows that the sentiment in analyst report texts positively correlates with stock returns on the report publication date.

I assign a sentiment score s to each firm-announcement by calculating the net fraction of positive words in associated report and then taking the average over the reports,

$$s = \frac{1}{N}\sum_{\text{reports}} \frac{\textit{positive words} - \textit{negative words}}{\textit{total number of words}}$$

Words are classified as positive, negative, or neutral using the Loughran and McDonald (2011) sentiment dictionary for financial data. Words are lemmatized before the classification. To form return predictions, I regress the announcement returns on the sentiment scores and a constant,

$$\arg\min_{b_0,b_1} (\mathbf{r} - [\mathbf{s}\ I]b)^\top \Omega(\mathbf{r} - [\mathbf{s}\ I]b),$$

where $\mathbf{s}$ is a vector of sentiment scores s, 1 is a vector of ones, and $[\mathbf{s}\ I]$ is the concatenation of the two. Ω is a diagonal matrix of sample weights, and is identical to the weight matrix used for the HAN model.

B.2.2 Elastic net

The elastic net solves the following equation

$$\arg\min_{b} \left((\mathbf{r} - \hat{\mathbf{r}})^\top \Omega(\mathbf{r} - \hat{\mathbf{r}}) + \lambda_1 \|b\|_1 + \lambda_2 \|b\|_2^2 \right)$$

where $\hat{\mathbf{r}}$ is a vector of firm-announcement retun predictions, with each element $\hat{r}$ being equal to the average prediction formed from each associated report,

$$\hat{r} = \frac{1}{N}\sum_{\text{reports}} x^\top b,$$

where x is a vector of word occurrence frequencies. Ω is a diagonal matrix of sample weights, and is identical to the weight matrix used for the HAN model.

The regularization parameters λ_1 and λ_2 are hyperparameters. The elastic net has three noteworthy special cases: if $\lambda_1 = \lambda_2 = 0$, it is identical to weighted least squares. If $\lambda_1 = 0$, it is equal to Ridge regression, and if $\lambda_2 = 0$ it is equal to LASSO. λ_2 penalizes large absolute coefficients, thus forcing the coefficients to be closer to zero. λ_1 tends to induce sparsity to the coefficient vector b. The idea of both the L1 and L2 regularization is to reduce the variance of the least-squares estimator, which often reduces overfitting and results in better out-of-sample predictions.

I use grid search to determine the values for λ_1 and λ_2 that minimize the loss function in the validation set. To ensure a level playing field with the HAN algorithm, I only choose the hyperparameters once for the 2004 data and hold them constant for all other training periods. The grid search procedure determines $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$ to be the optimal parameters.

B.3 Robustness checks

B.3.1 Monthly portfolio sorts excluding non-announcing firms

	Low	2	3	4	High	High-Low
Intercept	-0.753***	-0.045	-0.104	0.343**	0.462***	1.214***
	(-2.63)	(-0.18)	(-0.53)	(2.00)	(2.83)	(4.08)
Mkt-RF	1.445***	1.360***	1.343***	1.217***	1.206***	-0.239***
	(20.57)	(21.57)	(27.68)	(28.91)	(30.15)	(-3.28)
N	192	192	192	192	192	192
$Adj.R^2$	0.69	0.71	0.80	0.81	0.83	0.05

Panel A: CAPM

	Low	2	3	4	High	High-Low
Intercept	-0.322	0.268	0.108	0.442***	0.493***	0.815***
	(-1.32)	(1.22)	(0.71)	(3.31)	(3.56)	(2.80)
Mkt-RF	1.102***	1.060***	1.080***	1.026***	1.060***	-0.042
	(15.66)	(16.75)	(24.63)	(26.71)	(26.66)	(-0.50)
SMB	0.814***	0.814***	0.707***	0.657***	0.615***	-0.199
	(7.09)	(7.90)	(9.90)	(10.49)	(9.49)	(-1.45)
HML	0.041	0.226**	0.317***	0.316***	0.038	-0.004
	(0.35)	(2.10)	(4.26)	(4.84)	(0.56)	(-0.03)
CMA	-0.099	-0.160	-0.184	-0.146	-0.142	-0.043
	(-0.52)	(-0.93)	(-1.54)	(-1.39)	(-1.31)	(-0.19)
RMW	-0.716***	-0.379**	-0.050	0.151*	0.211**	0.927***
	(-4.31)	(-2.54)	(-0.48)	(1.67)	(2.24)	(4.68)
Mom	-0.137**	-0.088	-0.154***	-0.075**	-0.113***	0.024
	(-2.25)	(-1.61)	(-4.06)	(-2.26)	(-3.29)	(0.33)
N	192	192	192	192	192	192
$Adj.R^2$	0.79	0.80	0.89	0.90	0.88	0.16

Panel B: FF6

The table shows regression results for equally-weighted portfolios sorted on the neural return prediction. At the end of each month, firms that announce earnings in the following month are sorted into quintiles based on their neural return prediction. *Low* denotes the lowest predicted return quintile, and *high* denotes the highest predicted return quintile. The sample period is January 2004 to December 2019. Returns are monthly and in percent. T-statistics are shown in parentheses.

Table B.1: Portfolios of announcing firms sorted on neural return prediction

B.3.2 Announcement and publication drifts for different firm sizes

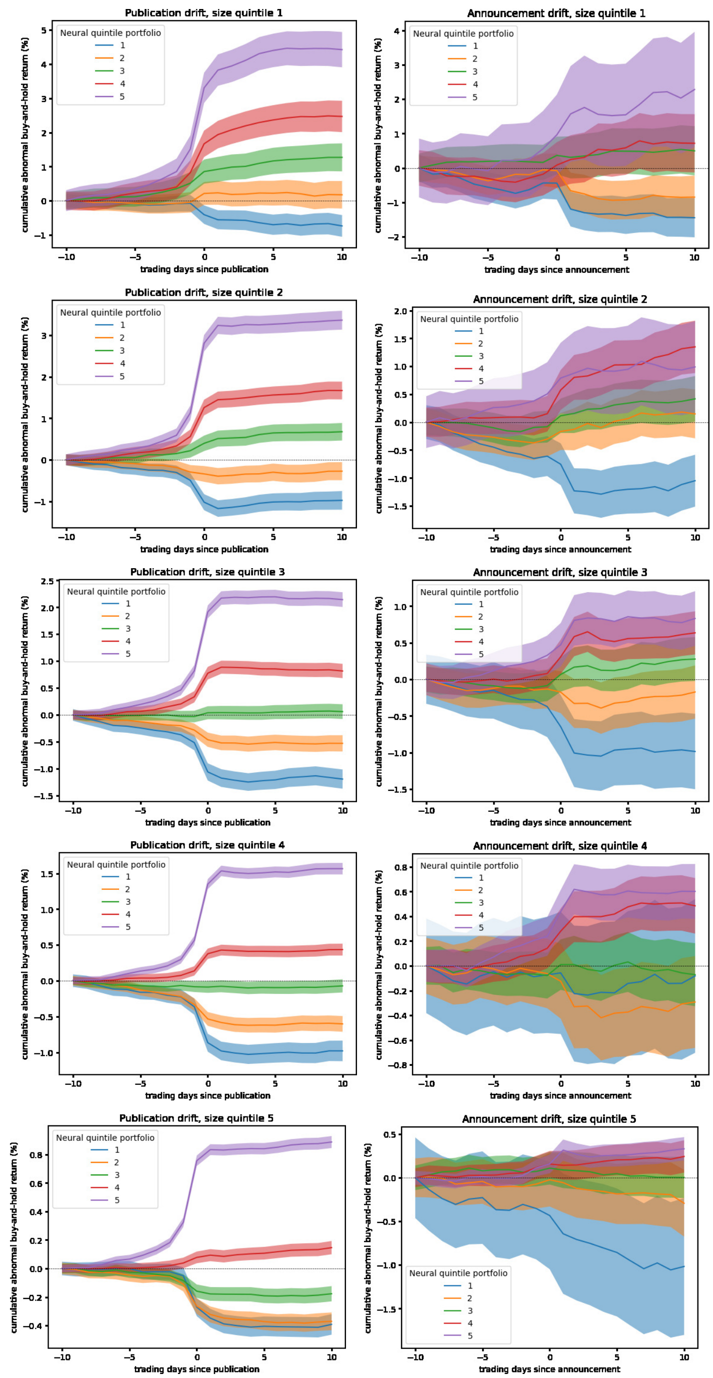


Figure 3.5 for different firm sizes. Firms are sorted into five size quintiles based on CRSP breakpoints three months prior to the publication.

Figure B.1: Publication and announcement drifts for different firm sizes

B.3.3 Controlling for characteristics

To test the relationship between the neural return predictions and common return predictors, I regress the realized announcement returns on the predicted returns and various control variables. In Table B.2 column (5), I control for various firm characteristics including market beta, size, profitability, investments, and momentum. Out of the five characteristics, only profitability and investments are statistically significant at the 1% and 5% levels, respectively. The finding is in line with Engelberg et al. (2018), who find that anomalies have excessively high returns on corporate news days. The coefficient on the NRP drops from 0.70 to 0.59, suggesting that a small fraction of the predictability can be explained by profitability and investments.

	(1)	(2)	(3)	(4)	(5)
Neural return prediction	0.841***	0.786***	0.704***		0.602***
	(18.28)	(15.13)	(12.76)		(9.61)
Publ. return				-0.001	-0.005
				(-0.22)	(-0.95)
Beta				0.015	0.025
				(0.39)	(0.67)
Size				0.056***	0.009
				(4.22)	(0.60)
B/M				0.030	0.018
				(0.54)	(0.32)
Profitability				0.467***	0.278***
				(6.11)	(3.67)
Investments				-0.168**	-0.139*
				(-2.27)	(-1.90)
Momentum				0.098**	0.082*
				(1.98)	(1.66)
Constant	-0.068***	-0.060**			
	(-3.37)	(-2.44)			
Fixed Effects			(t)	(t)	(t)
Sample	All	Controls	Controls	Controls	Controls
Observations	170,028	92,380	92,295	92,295	92,295
R^2	0.0048	0.0037	0.0544	0.0533	0.0547
Within R^2	0.0048	0.0037	0.0027	0.0016	0.0030

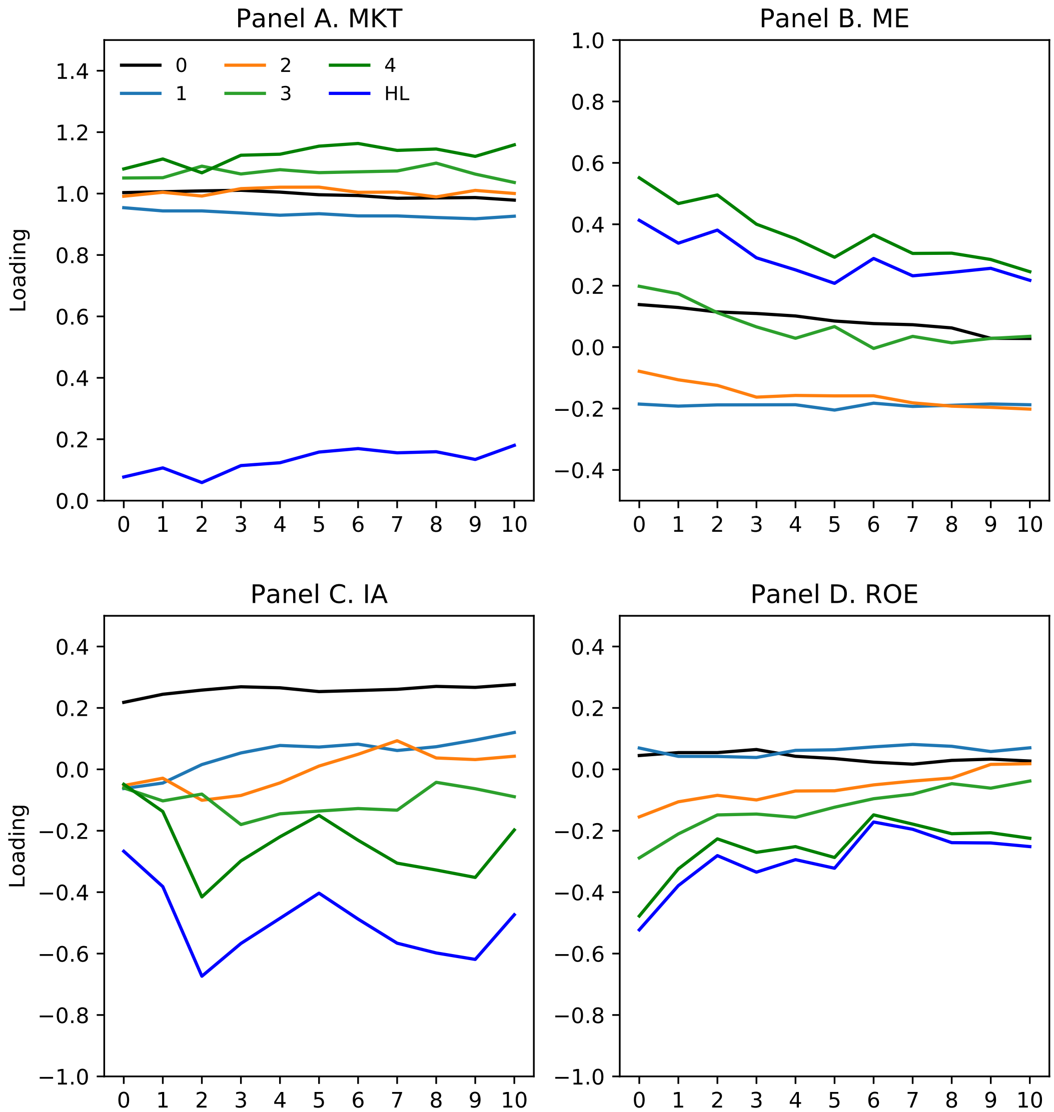
The table shows regressions of cumulative announcement returns on out-of-sample predictions. The dependent variable is the market-adjusted return in the two-day announcement window. Column (1) shows results for the sample of all announcements with non-missing neural return prediction. In columns (2)-(5) the sample is restricted to observations for which all control variables are available. Columns (3)-(5) use daily date fixed effects as indicated in the fixed effects row. Since every announcement reaction is measured over the span of two trading days and the two-day windows of different firms might only partially overlap, every observation has two date-fixed effects corresponding to the two days in the return measurement window. Standard errors are triple-clustered by firm and the two dates. t-Statistics are shown in parentheses.

Table B.2: Out-of-sample announcement return regressions controlling for characteristics

Appendix C

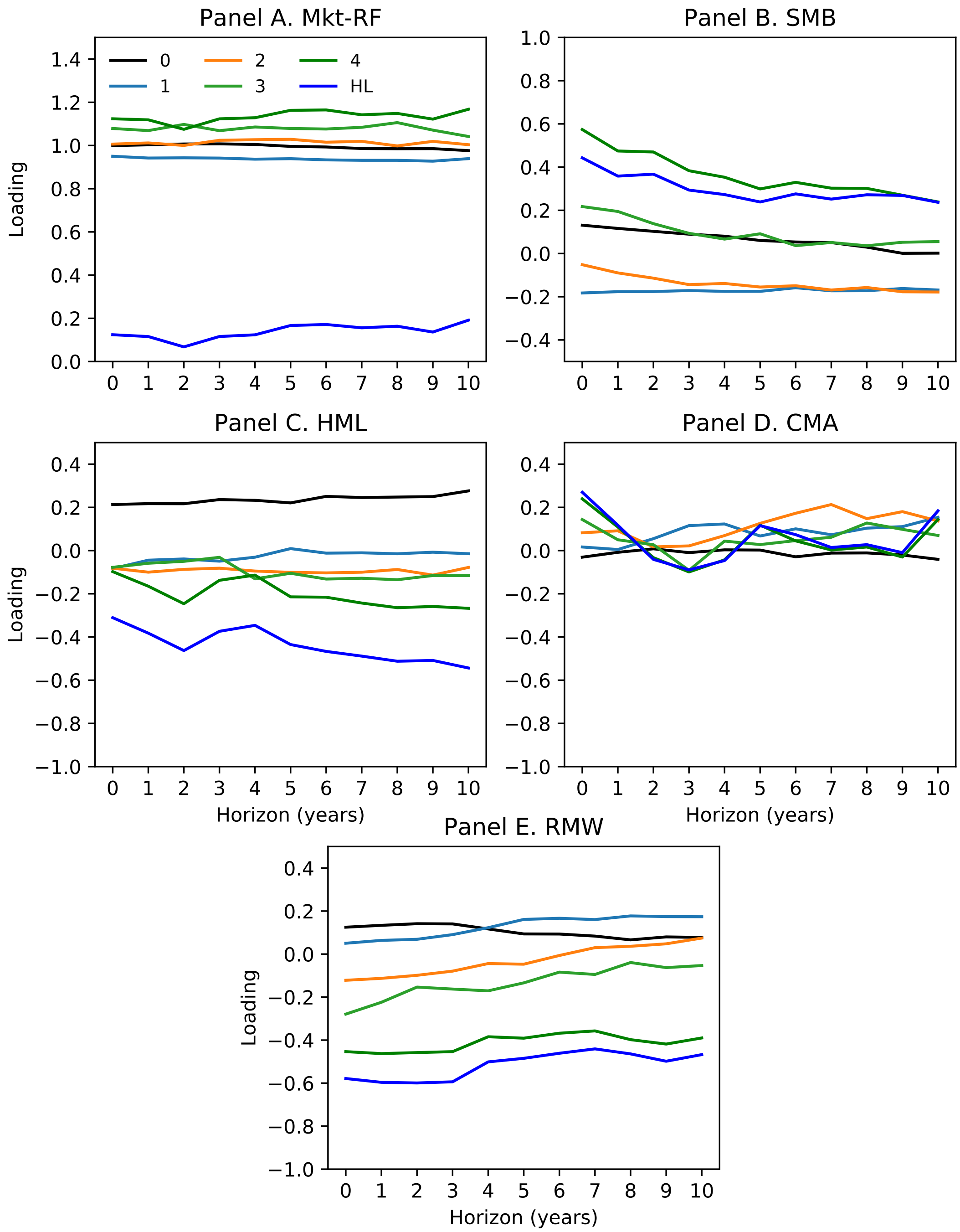
Appendix to Chapter 4

C.1 Additional Tables and Figures



The figure shows the dynamics of loadings of PI-sorted aged portfolios on the Q4 factors as indicated in headings of panels A-D. The construction of the underlying portfolios is described in detail in notes to Table 4.8. The time period of the sample is given by availability of the Q-factors, i.e., 1967-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Figure C.1: Aged Patent Intensity portfolios, q4-factor loading dynamics, 1967-2021



The figure shows the dynamics of FF5 factor loadings of PI-sorted aged portfolios. Refer to Fig. C.1 for more details.

Figure C.2: Aged Patent Intensity portfolios, FF5-factor loading dynamics, 1963-2021

	0	1	2	HL
Panel A. Excess returns				
Excess return	9.4**	10.69***	14.01***	4.61***
	(2.48)	(2.85)	(2.86)	(2.61)
Panel B. CAPM				
Constant	-0.64	0.75*	1.3	1.93
	(-0.97)	(1.7)	(1.14)	(1.4)
Mkt-RF	0.98***	0.97***	1.23***	0.26***
	(44.55)	(55.46)	(31.99)	(7.79)
R^2	0.96	0.98	0.95	0.36
Panel C. Fama-French 1993				
Constant	-0.85	0.9**	1.12	1.97
	(-1.34)	(2.04)	(1.01)	(1.43)
Mkt-RF	0.91***	1.01***	1.18***	0.27***
	(63.99)	(86.58)	(46.49)	(9.01)
SMB	0.08***	-0.08***	0.07*	-0.01
	(2.63)	(-3.99)	(1.68)	(-0.29)
HML	0.13***	-0.08***	0.1*	-0.03
	(4.55)	(-3.37)	(1.81)	(-0.45)
R^2	0.97	0.98	0.95	0.36

The table shows the average excess returns of PI-sorted portfolios in panel A and results of regressing the portfolio returns on a constant and market excess returns, and Fama-French 3 factors in panel B and C, respectively. Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios each year at the end of June. All portfolios are value-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, but the estimates of the average excess returns and constants are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is 1926-1963. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.1: Returns of PI-sorted portfolios, 1926-1963

	1926-2021				1963-2021					
	0	1	2	HL	0	1	2	3	4	HL
Panel A. Excess returns										
Excess return	10.64***	10.14***	14.89***	4.25***	8.79***	7.85***	9.25***	11.07***	15.72***	6.92***
	(3.83)	(4.28)	(4.55)	(3.92)	(3.06)	(3.33)	(3.39)	(3.43)	(3.98)	(3.74)
Panel B. CAPM										
Constant	1.1	0.66	3.69**	2.58***	1.7	0.19	1.1	2.56	7.09***	5.38***
	(0.95)	(1.09)	(2.56)	(2.61)	(1.08)	(0.29)	(0.96)	(1.45)	(2.87)	(3.17)
Mkt-RF	1.15***	1.15***	1.35***	0.2***	1.01***	1.1***	1.17***	1.22***	1.23***	0.22***
	(24.4)	(59.19)	(29.69)	(12.01)	(30.69)	(88.15)	(44.11)	(36.46)	(26.04)	(7.28)
R^2	0.78	0.93	0.76	0.15	0.71	0.92	0.81	0.68	0.52	0.07
Panel C. Fama-French 1993										
Constant	-0.77*	0.21	2.16**	2.94***	-0.6	-0.02	0.24	1.2	5.13***	5.73***
	(-1.72)	(0.55)	(2.36)	(3.02)	(-0.98)	(-0.03)	(0.36)	(1.13)	(3.09)	(3.44)
Mkt-RF	0.92***	1.06***	1.11***	0.19***	0.88***	1.01***	1.01***	1.0***	0.95***	0.07
	(79.54)	(60.6)	(42.42)	(8.64)	(62.36)	(84.57)	(43.78)	(30.7)	(22.78)	(1.58)
SMB	0.86***	0.43***	1.1***	0.24**	0.88***	0.37***	0.75***	1.1***	1.45***	0.57***
	(39.87)	(9.41)	(11.68)	(2.28)	(23.52)	(14.99)	(16.88)	(12.26)	(10.83)	(3.71)
HML	0.41***	0.02	0.19***	-0.21***	0.32***	-0.05***	0.0	0.03	0.08	-0.24***
	(14.08)	(0.75)	(3.09)	(-3.69)	(12.37)	(-3.0)	(0.06)	(0.66)	(1.0)	(-2.8)
R^2	0.97	0.97	0.93	0.28	0.95	0.97	0.95	0.92	0.82	0.34
Panel D. Fama-French 2015										
Constant					-0.45	0.2	1.08*	2.57**	7.21***	7.66***
					(-0.74)	(0.45)	(1.94)	(2.46)	(4.2)	(4.36)
Mkt-RF					0.87***	1.02***	1.01***	0.99***	0.94***	0.07**
					(63.36)	(91.5)	(59.37)	(46.1)	(27.25)	(2.07)
SMB					0.88***	0.34***	0.69***	0.98***	1.27***	0.39***
					(29.9)	(22.3)	(26.47)	(21.83)	(18.66)	(5.45)
HML					0.16***	-0.15***	-0.15***	-0.2***	-0.25***	-0.41***
					(5.2)	(-8.25)	(-4.69)	(-3.29)	(-3.03)	(-4.66)
CMA					0.06	0.11***	0.13**	0.2**	0.33***	0.27**
					(1.49)	(3.01)	(2.22)	(2.41)	(2.69)	(2.37)
RMW					-0.03	-0.11***	-0.26***	-0.42***	-0.67***	-0.63***
					(-1.13)	(-4.42)	(-4.72)	(-4.39)	(-4.73)	(-4.02)
R^2					0.95	0.97	0.96	0.93	0.85	0.46

The table shows the average excess returns of PI-sorted, equal-weighted portfolios in panel A and results of regressing the portfolio returns on a constant and market excess returns, Fama-French 3 factors and Fama-French 5 factors in panels B, C, and D, respectively. Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios each year at the end of June. All portfolios are equal-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, but the estimates of the average excess returns and constants are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is indicated in headings, i.e., 1926-2021 and 1963-2021. Data for Fama-French 5 factors is available only from 1963. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.2: Returns of PI-sorted (equal-weighted) portfolios

	0	1	2	3	4	HL
Constant	-1.58***	0.03	2.35***	2.85***	5.73***	7.31***
	(-3.38)	(0.06)	(3.33)	(2.71)	(3.58)	(4.15)
Mkt-RF	1.0***	0.95***	1.0***	1.06***	1.11***	0.11**
	(81.01)	(93.2)	(66.46)	(36.24)	(26.05)	(2.35)
SMB	0.13***	-0.18***	-0.05*	0.22***	0.58***	0.45***
	(5.72)	(-14.08)	(-1.66)	(3.9)	(7.4)	(4.74)
HML	0.2***	-0.08***	-0.1**	-0.13*	-0.15	-0.35***
	(6.85)	(-3.33)	(-2.29)	(-1.9)	(-1.57)	(-3.16)
CMA	-0.02	0.01	0.09	0.18**	0.27*	0.29*
	(-0.62)	(0.43)	(1.56)	(2.01)	(1.8)	(1.74)
RMW	0.13***	0.05**	-0.12***	-0.26***	-0.43***	-0.56***
	(3.38)	(2.55)	(-3.37)	(-3.14)	(-2.77)	(-3.15)
Mom	-0.02	0.01	-0.03	-0.1***	-0.09	-0.07
	(-1.33)	(0.95)	(-1.53)	(-2.84)	(-1.5)	(-1.0)
R^2	0.96	0.96	0.9	0.86	0.78	0.32

The table shows the results of regressing the portfolio returns on a constant, Fama-French five factors and momentum factor. Portfolio "0" consists of non-patenting firms and the remaining portfolios of patenting firms sorted by *PI*. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio "0". Stocks are sorted into portfolios each year at the end of June. All portfolios are value-weighted and rebalanced annually. The underlying portfolio returns are at monthly frequency, but the estimates of the constant are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is 1963-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.3: Patent Intensity sorts and performance, Fama-French Factors with Momentum, 1963-2021

Horizon (years)	0	1	2	3	4	HL
0	-1.58***	0.03	2.35***	2.85***	5.73***	7.31***
	(-3.38)	(0.06)	(3.33)	(2.71)	(3.58)	(4.15)
1	-1.61***	0.03	2.5***	1.63*	6.32***	7.93***
	(-3.48)	(0.07)	(3.64)	(1.73)	(3.67)	(4.15)
2	-1.56***	0.17	1.58**	1.38	4.67***	6.23***
	(-3.34)	(0.38)	(2.38)	(1.58)	(2.75)	(3.28)
3	-1.77***	-0.06	1.81**	1.96**	3.09**	4.86***
	(-3.87)	(-0.14)	(2.54)	(2.11)	(2.23)	(3.05)
4	-1.7***	-0.02	1.13*	2.39***	2.32*	4.02***
	(-3.71)	(-0.05)	(1.8)	(2.65)	(1.86)	(2.72)
5	-1.6***	0.27	0.77	2.25**	2.77**	4.37***
	(-3.48)	(0.56)	(1.23)	(2.46)	(2.1)	(2.93)
6	-1.47***	0.15	1.03*	1.51*	3.17**	4.64***
	(-3.37)	(0.34)	(1.76)	(1.67)	(2.18)	(2.85)
7	-1.29***	0.15	0.55	2.01**	2.98**	4.27***
	(-2.93)	(0.35)	(0.94)	(2.39)	(2.12)	(2.71)
8	-1.51***	-0.03	1.38**	0.8	3.27**	4.78***
	(-3.39)	(-0.07)	(2.24)	(0.85)	(2.35)	(3.01)
9	-1.47***	0.31	0.64	1.22	3.89***	5.36***
	(-3.17)	(0.7)	(1.1)	(1.46)	(2.84)	(3.39)
10	-1.22***	0.2	-0.15	2.33**	1.66	2.88*
	(-2.64)	(0.43)	(-0.24)	(2.52)	(1.12)	(1.7)

The table shows the abnormal returns (alphas) relative to FF5 model with momentum (Fama and French 2015) of PI-sorted portfolios for holding period of one-year at different investment horizons (indicated in rows). Portfolio "0" consists of non-patenting firms and the remaining portfolios of patenting firms sorted by PI. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio "0". Stocks are sorted into portfolios at the end of June K years prior to the beginning of the holding period in July of year t. The holding period lasts for one year from July (end of June) in year t to the end of June in year $t+1$. Each portfolio consists of the stocks assigned to the portfolio K years ago that are still active as of the beginning of the holding period, i.e., end of June in year t. Portfolios are value weighted with weights as of the beginning of the holding period. The underlying portfolio returns are at monthly frequency, but the estimates of the alphas are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the FF5-factors, i.e., 1963-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.4: Aged Patent Intensity portfolios, FF5+Momentum alpha dynamics, 1963-2021

Horizon (years)	0	1	2	3	4	HL
0	-2.04***	-0.04	2.83***	3.5***	6.79***	8.82***
	(-3.27)	(-0.08)	(3.75)	(2.87)	(3.69)	(3.96)
1	-2.06***	0.05	2.78***	2.29**	6.77***	8.83***
	(-3.44)	(0.11)	(3.69)	(2.07)	(3.44)	(3.82)
2	-1.99***	0.17	1.89**	1.59*	5.26***	7.25***
	(-3.21)	(0.37)	(2.57)	(1.65)	(2.63)	(3.05)
3	-2.4***	0.14	2.17***	2.01**	2.73*	5.13***
	(-3.79)	(0.29)	(2.62)	(2.01)	(1.84)	(2.85)
4	-2.19***	0.06	1.55**	2.45**	2.18	4.37***
	(-3.76)	(0.12)	(2.19)	(2.45)	(1.56)	(2.6)
5	-2.17***	0.33	1.12	2.32**	2.5*	4.67***
	(-3.94)	(0.6)	(1.61)	(2.27)	(1.65)	(2.61)
6	-2.1***	0.52	1.11*	1.45	2.81*	4.91**
	(-3.42)	(0.99)	(1.69)	(1.43)	(1.67)	(2.41)
7	-1.77***	0.22	0.93	1.89*	2.83*	4.6**
	(-3.06)	(0.47)	(1.5)	(1.93)	(1.74)	(2.36)
8	-2.09***	0.31	1.42**	0.59	2.74*	4.83**
	(-3.78)	(0.62)	(2.2)	(0.55)	(1.7)	(2.5)
9	-2.11***	0.72	0.71	0.84	3.51**	5.62***
	(-3.51)	(1.37)	(1.14)	(0.91)	(2.04)	(2.71)
10	-1.81***	0.53	-0.21	2.03*	1.5	3.32
	(-2.95)	(1.0)	(-0.31)	(1.96)	(0.85)	(1.55)

The table shows the abnormal returns (alphas) relative to five-factor model (Fama and French 2015) of PI-sorted portfolios for holding period of one-year at different investment horizons (indicated in rows). Portfolio ”0” consists of non-patenting firms and the remaining portfolios of patenting firms sorted by PI. HL is a zero-cost portfolio with a long position in the highest PI portfolio and a short position in portfolio ”0”. Stocks are sorted into portfolios at the end of June K years prior to the beginning of the holding period in July of year t. The holding period lasts for one year from July (end of June) in year t to the end of June in year $t+1$. Each portfolio consists of the stocks assigned to the portfolio K years ago that are still active as of the beginning of the holding period, i.e., end of June in year t. Portfolios are value weighted with weights as of the beginning of the holding period. The underlying portfolio returns are at monthly frequency, but the estimates of the alphas are annualized. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the Q-factors, i.e., 1967-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.5: Aged Patent Intensity portfolios, q4 alpha dynamics, 1967-2021

		Ex. Ret.	Fama-French 2015						
		Constant	Constant	Mkt-RF	SMB	HML	CMA	RMW	R^2
Beta	Non-Inno	-0.841	-1.713	0.641***	-0.039	0.118	-0.819***	-0.315***	0.43
		(-0.33)	(-0.83)	(15.04)	(-0.64)	(1.43)	(-6.69)	(-3.73)	
	Inno	-2.077	-1.069	0.576***	-0.371***	-0.099	-0.639***	-0.514***	0.33
		(-0.77)	(-0.45)	(11.74)	(-5.28)	(-1.05)	(-4.54)	(-5.28)	
	Diff	-1.236	0.644	-0.066	-0.332***	-0.217***	0.179	-0.199**	0.07
		(-0.66)	(0.33)	(-1.65)	(-5.82)	(-2.83)	(1.57)	(-2.52)	
Size	Non-Inno	-4.541*	-3.553*	0.200***	-1.131***	-0.301***	0.157	0.183**	0.41
		(-1.87)	(-1.81)	(5.02)	(-19.37)	(-3.98)	(1.35)	(2.26)	
	Inno	-5.517*	-4.409*	0.101**	-1.339***	-0.441***	0.134	0.673***	0.46
		(-1.77)	(-1.83)	(2.07)	(-18.70)	(-4.76)	(0.95)	(6.80)	
	Diff	-0.976	-0.856	-0.099***	-0.208***	-0.140**	-0.023	0.490***	0.18
		(-0.57)	(-0.52)	(-2.95)	(-4.24)	(-2.21)	(-0.23)	(7.24)	
B/M	Non-Inno	3.588*	-0.529	0.019	0.156***	1.131***	0.061	0.006	0.57
		(1.82)	(-0.39)	(0.70)	(3.89)	(21.73)	(0.76)	(0.10)	
	Inno	1.485	-3.142*	0.173***	0.298***	1.151***	0.104	-0.354***	0.44
		(0.61)	(-1.65)	(4.48)	(5.28)	(15.74)	(0.93)	(-4.53)	
	Diff	-2.104	-2.614	0.153***	0.142**	0.020	0.043	-0.359***	0.08
		(-0.94)	(-1.15)	(3.34)	(2.12)	(0.24)	(0.32)	(-3.87)	
Invest	Non-Inno	-3.607***	-1.192	0.039*	-0.138***	-0.068	-0.839***	0.158***	0.41
		(-2.69)	(-1.09)	(1.77)	(-4.25)	(-1.62)	(-13.11)	(3.53)	
	Inno	-2.378	2.944**	-0.038	-0.190***	-0.001	-1.433***	-0.070	0.47
		(-1.32)	(2.12)	(-1.36)	(-4.60)	(-0.03)	(-17.56)	(-1.23)	
	Diff	1.229	4.136**	-0.077**	-0.052	0.066	-0.594***	-0.228***	0.06
		(0.68)	(2.23)	(-2.06)	(-0.95)	(0.93)	(-5.45)	(-3.00)	
Profit	Non-Inno	5.653**	1.847	-0.113***	-0.203***	-0.112	0.226*	1.541***	0.46
		(2.17)	(0.92)	(-2.78)	(-3.42)	(-1.46)	(1.92)	(18.75)	
	Inno	-2.202	-3.421*	-0.145***	-0.691***	0.184**	-0.344***	1.478***	0.58
		(-0.76)	(-1.75)	(-3.69)	(-11.97)	(2.47)	(-3.01)	(18.52)	
	Diff	-7.854***	-5.268*	-0.033	-0.487***	0.297***	-0.570***	-0.063	0.07
		(-2.76)	(-1.83)	(-0.56)	(-5.75)	(2.70)	(-3.39)	(-0.54)	

The table shows the average excess returns of innovative and non-innovative firms sorted on common firm characteristics as well as the results of regressing the portfolio returns on a constant and Fama-French 5 factors. Stocks are labeled as innovators and non-innovators at the end of June in each year and independently sorted into five portfolios. Innovative firms are firms that have at least three patents over the last three years and one patent over the last year at the time of portfolio formation. The table shows the returns of a long-short portfolio that goes long the highest quintile and short the lowest quintile. More details can be found in the caption of Table 4.11.

Table C.6: Characteristics-sorted portfolios in innovative vs. non-innovative firms with full-sample breakpoints

		Ex. Ret.	q5						
		Constant	Constant	MKT	ME	IA	ROE	EG	R^2
Beta	Non-Inno	-1.493	0.922	0.665***	-0.121*	-0.353***	-0.012	-0.498***	0.43
		(-0.55)	-0.38	-14.71	(-1.96)	(-3.45)	(-0.15)	(-4.11)	
	Inno	-2.902	-1.092	0.643***	-0.416***	-0.454***	-0.237**	-0.132	0.31
		(-1.00)	(-0.38)	-12.1	(-5.72)	(-3.77)	(-2.46)	(-0.93)	
	Diff	-1.409	-2.014	-0.022	-0.295***	-0.101	-0.225***	0.367***	0.08
		(-0.71)	(-0.89)	(-0.52)	(-5.13)	(-1.06)	(-2.96)	-3.26	
Size	Non-Inno	-4.421*	-7.848***	0.238***	-0.875***	0.103	0.677***	0.009	0.41
		(-1.73)	(-3.40)	-5.55	(-14.73)	-1.07	-8.69	-0.08	
	Inno	-5.799*	-10.452***	0.177***	-1.181***	0.048	1.050***	0.071	0.51
		(-1.75)	(-3.84)	-3.51	(-16.90)	-0.42	-11.45	-0.53	
	Diff	-1.377	-2.604	-0.061*	-0.306***	-0.055	0.372***	0.063	0.19
		(-0.75)	(-1.35)	(-1.69)	(-6.15)	(-0.68)	-5.71	-0.66	
B/M	Non-Inno	3.920*	0.927	-0.009	0.121**	1.172***	-0.504***	0.104	0.34
		(1.89)	-0.47	(-0.25)	-2.38	-14.26	(-7.58)	-1.07	
	Inno	1.684	-0.223	0.124***	0.268***	1.136***	-0.842***	0.073	0.35
		(0.66)	(-0.09)	-2.74	-4.29	-11.19	(-10.24)	-0.61	
	Diff	-2.236	-1.15	0.133***	0.148**	-0.036	-0.337***	-0.031	0.08
		(-0.95)	(-0.43)	-2.7	-2.16	(-0.33)	(-3.76)	(-0.24)	
Invest	Non-Inno	-4.045***	-0.35	0.059**	-0.211***	-0.769***	0.266***	-0.201***	0.38
		(-2.93)	(-0.27)	-2.49	(-6.44)	(-14.46)	-6.19	(-3.21)	
	Inno	-2.875	0.831	0.045	-0.198***	-1.391***	0.097*	0.169**	0.43
		(-1.51)	-0.49	-1.43	(-4.56)	(-19.74)	-1.71	-2.04	
	Diff	1.170	1.181	-0.014	0.013	-0.623***	-0.169**	0.371***	0.08
		(0.62)	-0.55	(-0.35)	-0.24	(-6.98)	(-2.34)	-3.52	
Profit	Non-Inno	5.432**	-1.228	-0.095**	-0.256***	0.261**	1.081***	0.056	0.37
		(2.04)	(-0.49)	(-2.05)	(-4.00)	-2.51	-12.88	-0.46	
	Inno	-1.735	-3.424	-0.125**	-0.829***	0.219**	0.950***	-0.144	0.43
		(-0.58)	(-1.28)	(-2.53)	(-12.09)	-1.97	-10.56	(-1.10)	
	Diff	-7.168**	-2.196	-0.031	-0.574***	-0.042	-0.131	-0.2	0.07
		(-2.47)	(-0.67)	(-0.50)	(-6.77)	(-0.30)	(-1.17)	(-1.23)	

The table shows the average excess returns of innovative and non-innovative firms sorted on common firm characteristics as well as the results of regressing the portfolio returns on a constant and q5-factors. Stocks are labeled as innovators and non-innovators at the end of June in each year and independently sorted into five portfolios. Innovative firms are firms that have at least three patents over the last three years and one patent over the last year at the time of portfolio formation. The table shows the returns of a long-short portfolio that goes long the highest quintile and short the lowest quintile. More details can be found in the caption of Table 4.11.

Table C.7: Characteristics-sorted portfolios in innovative vs. non-innovative firms with the q5-Factor model and full-sample breakpoints

		Ex. Ret.	q4					
		Constant	Constant	MKT	ME	IA	ROE	R^2
Beta	Non-Inno	-1.941	-3.038	0.718***	-0.075	-0.454***	-0.163**	0.41
		(-0.73)	(-1.35)	(16.39)	(-1.22)	(-4.52)	(-2.19)	
	Inno	-0.706	-2.139	0.657***	-0.404***	-0.480***	-0.277***	0.31
		(-0.23)	(-0.82)	(12.92)	(-5.64)	(-4.12)	(-3.21)	
	Diff	1.235	0.899	-0.061	-0.329***	-0.027	-0.114*	0.06
		-0.52	(0.43)	(-1.51)	(-5.77)	(-0.29)	(-1.66)	
Size	Non-Inno	-4.236	-7.777***	0.237***	-0.876***	0.104	0.680***	0.41
		(-1.61)	(-3.67)	(5.82)	(-15.01)	(1.11)	(9.89)	
	Inno	-7.544**	-9.875***	0.169***	-1.188***	0.061	1.073***	0.51
		(-2.46)	(-3.96)	(3.52)	(-17.29)	(0.55)	(13.26)	
	Diff	-3.308*	-2.098	-0.068**	-0.312***	-0.043	0.393***	0.19
		(-1.68)	(-1.18)	(-2.00)	(-6.38)	(-0.55)	(6.82)	
B/M	Non-Inno	4.283**	1.768	-0.021	0.111**	1.192***	-0.471***	0.34
		-2.33	(0.98)	(-0.61)	(2.22)	(14.86)	(-8.02)	
	Inno	5.290**	0.365	0.115***	0.261***	1.150***	-0.818***	0.35
		-2.28	(0.16)	(2.69)	(4.25)	(11.62)	(-11.29)	
	Diff	1.006	-1.403	0.137***	0.151**	-0.042	-0.347***	0.08
		-0.49	(-0.58)	(2.92)	(2.24)	(-0.39)	(-4.39)	
Invest	Non-Inno	-4.556***	-1.975*	0.083***	-0.192***	-0.806***	0.202***	0.37
		(-3.74)	(-1.68)	(3.65)	(-5.92)	(-15.44)	(5.27)	
	Inno	-2.679	2.200	0.025	-0.214***	-1.360***	0.152***	0.43
		(-1.30)	(1.41)	(0.84)	(-5.00)	(-19.73)	(3.01)	
	Diff	1.877	4.175**	-0.058	-0.022	-0.554***	-0.049	0.06
		-0.96	(2.11)	(-1.51)	(-0.41)	(-6.30)	(-0.77)	
Profit	Non-Inno	6.225**	-0.774	-0.101**	-0.261***	0.271***	1.099***	0.37
		-2.35	(-0.34)	(-2.31)	(-4.15)	(2.68)	(14.85)	
	Inno	-3.65	-4.586*	-0.108**	-0.816***	0.192*	0.904***	0.43
		(-1.20)	(-1.87)	(-2.31)	(-12.09)	(1.77)	(11.38)	
	Diff	-9.876***	-3.812	-0.007	-0.555***	-0.079	-0.195**	0.07
		(-3.44)	(-1.26)	(-0.12)	(-6.65)	(-0.59)	(-1.99)	

The table shows the average excess returns of innovative and non-innovative firms sorted on common firm characteristics as well as the results of regressing the portfolio returns on a constant and q4-factors. Stocks are labeled as innovators and non-innovators at the end of June in each year and independently sorted into five portfolios. Innovative firms are firms that have at least three patents over the last three years and one patent over the last year at the time of portfolio formation. The table shows the returns of a long-short portfolio that goes long the highest quintile and short the lowest quintile. More details can be found in the caption of Table 4.11.

Table C.8: Characteristics-sorted portfolios in innovative vs. non-innovative firms with the q4-Factor model and full-sample breakpoints

	Panel A. MKT					
Horizon (years)	0	1	2	3	4	HL
0	0.98***	0.95***	1.02***	1.08***	1.15***	0.17***
	(55.85)	(86.26)	(60.46)	(29.18)	(23.18)	(2.85)
1	0.99***	0.94***	1.03***	1.08***	1.16***	0.17***
	(55.89)	(79.07)	(59.05)	(36.91)	(31.16)	(3.78)
2	0.99***	0.95***	1.02***	1.11***	1.1***	0.11**
	(57.22)	(69.85)	(52.34)	(47.16)	(23.82)	(1.97)
3	0.99***	0.94***	1.04***	1.08***	1.15***	0.16***
	(57.88)	(73.03)	(51.96)	(47.28)	(28.67)	(3.24)
4	0.99***	0.94***	1.04***	1.09***	1.17***	0.18***
	(61.68)	(83.78)	(65.76)	(44.25)	(30.03)	(3.69)
5	0.98***	0.94***	1.03***	1.09***	1.19***	0.22***
	(59.66)	(75.58)	(60.55)	(45.74)	(26.81)	(4.01)
6	0.98***	0.94***	1.02***	1.08***	1.2***	0.22***
	(54.97)	(77.33)	(66.78)	(47.67)	(28.35)	(4.24)
7	0.97***	0.94***	1.02***	1.09***	1.17***	0.2***
	(61.09)	(91.42)	(63.97)	(53.25)	(27.45)	(3.78)
8	0.97***	0.94***	1.0***	1.11***	1.17***	0.2***
	(59.86)	(91.77)	(62.31)	(52.43)	(26.28)	(3.65)
9	0.97***	0.94***	1.02***	1.08***	1.14***	0.17***
	(57.63)	(88.71)	(64.0)	(44.41)	(22.36)	(2.81)
10	0.96***	0.95***	1.0***	1.05***	1.18***	0.21***
	(53.19)	(79.31)	(51.9)	(39.24)	(21.49)	(3.21)
10-0	-0.02**	0.0	-0.02	-0.04	0.02	0.04
	(-2.26)	(-0.25)	(-0.89)	(-1.13)	(0.57)	(0.95)

The table shows the loadings on the q-factors (indicated in panel headings) of PI-sorted portfolios for holding period of one-year at different investment horizons (indicated in rows). The last row shows the difference in loadings between horizon 10 and 0. The construction of the underlying portfolios is described in detail in notes to Table 4.8. Newey-West heteroscedasticity and autocorrelation consistent standard errors with five lags are reported in parentheses. The time period of the sample is given by availability of the q factors, i.e., 1967-2021. */**/*** indicate significance level at 10, 5, and 1%, respectively.

Table C.9: Aged Patent Intensity portfolios, q5-factor loadings dynamics, 1967-2021

Horizon (years)	Panel B. ME					
	0	1	2	3	4	HL
0	0.12**	-0.19***	-0.05	0.23***	0.61***	0.49***
	(2.36)	(-12.25)	(-1.44)	(2.99)	(5.12)	(2.98)
1	0.11**	-0.19***	-0.08**	0.2***	0.51***	0.39***
	(2.28)	(-14.31)	(-2.15)	(3.65)	(6.25)	(3.25)
2	0.1*	-0.19***	-0.1***	0.13***	0.52***	0.42***
	(1.95)	(-12.25)	(-3.58)	(3.01)	(4.41)	(2.62)
3	0.09*	-0.18***	-0.14***	0.08**	0.42***	0.33**
	(1.88)	(-13.09)	(-6.42)	(2.0)	(4.76)	(2.55)
4	0.09*	-0.18***	-0.14***	0.04	0.38***	0.3***
	(1.9)	(-11.5)	(-6.36)	(1.03)	(5.21)	(2.79)
5	0.07*	-0.2***	-0.15***	0.09**	0.33***	0.26**
	(1.68)	(-8.21)	(-5.57)	(2.04)	(3.76)	(2.11)
6	0.06	-0.17***	-0.15***	0.01	0.39***	0.33**
	(1.29)	(-8.4)	(-6.55)	(0.18)	(4.17)	(2.44)
7	0.06	-0.18***	-0.17***	0.05	0.33***	0.27*
	(1.36)	(-9.79)	(-6.84)	(1.29)	(3.1)	(1.88)
8	0.05	-0.18***	-0.19***	0.03	0.33***	0.28**
	(1.26)	(-8.51)	(-7.64)	(0.66)	(3.28)	(2.09)
9	0.02	-0.17***	-0.19***	0.04	0.3***	0.29**
	(0.37)	(-7.88)	(-7.26)	(0.87)	(2.95)	(2.05)
10	0.02	-0.17***	-0.2***	0.04	0.26**	0.25
	(0.36)	(-7.98)	(-8.31)	(0.94)	(2.33)	(1.61)
10-0	-0.1***	0.02	-0.15***	-0.18***	-0.35***	-0.25***
	(-8.05)	(0.64)	(-3.56)	(-2.93)	(-5.15)	(-3.35)

Table C.9-continued

Horizon (years)	Panel C. IA					
	0	1	2	3	4	HL
0	0.25***	-0.06**	-0.11***	-0.12*	-0.17	-0.42***
	(5.13)	(-2.15)	(-2.59)	(-1.7)	(-1.39)	(-2.8)
1	0.28***	-0.04	-0.08	-0.15**	-0.21	-0.49***
	(5.39)	(-1.5)	(-1.57)	(-2.23)	(-1.43)	(-2.6)
2	0.29***	0.01	-0.14**	-0.12*	-0.47***	-0.76***
	(5.74)	(0.47)	(-2.27)	(-1.84)	(-3.3)	(-4.21)
3	0.3***	0.04*	-0.12**	-0.21***	-0.35***	-0.65***
	(5.71)	(1.82)	(-2.09)	(-2.94)	(-3.93)	(-5.48)
4	0.3***	0.06**	-0.08*	-0.17***	-0.28***	-0.58***
	(6.15)	(2.41)	(-1.71)	(-3.21)	(-3.44)	(-5.4)
5	0.28***	0.06	-0.01	-0.17***	-0.22**	-0.5***
	(7.03)	(1.58)	(-0.23)	(-3.13)	(-2.32)	(-4.54)
6	0.28***	0.06	0.03	-0.15**	-0.29***	-0.57***
	(6.36)	(1.57)	(0.57)	(-2.31)	(-3.04)	(-4.68)
7	0.29***	0.04	0.07*	-0.16***	-0.35***	-0.64***
	(6.29)	(1.13)	(1.88)	(-3.05)	(-3.77)	(-5.25)
8	0.29***	0.05	0.03	-0.07	-0.37***	-0.67***
	(7.34)	(1.31)	(0.58)	(-1.14)	(-3.99)	(-5.71)
9	0.29***	0.07**	0.02	-0.08	-0.39***	-0.68***
	(6.97)	(1.99)	(0.43)	(-1.56)	(-4.03)	(-5.79)
10	0.3***	0.08***	0.04	-0.1	-0.23**	-0.53***
	(6.28)	(2.82)	(0.88)	(-1.58)	(-2.05)	(-4.07)
10-0	0.05**	0.14***	0.14**	0.01	-0.06	-0.11
	(2.11)	(3.8)	(2.42)	(0.16)	(-0.49)	(-0.8)

Table C.9-continued

Horizon (years)	Panel D. ROE					
	0	1	2	3	4	HL
0	0.1***	0.08***	-0.25***	-0.38***	-0.69***	-0.79***
	(2.7)	(3.31)	(-5.57)	(-5.35)	(-6.84)	(-6.37)
1	0.11***	0.05**	-0.19***	-0.3***	-0.45***	-0.56***
	(2.73)	(2.04)	(-4.98)	(-5.87)	(-4.59)	(-4.57)
2	0.11***	0.04	-0.15***	-0.21***	-0.32***	-0.43***
	(2.71)	(1.4)	(-3.51)	(-4.56)	(-2.95)	(-3.24)
3	0.12***	0.02	-0.17***	-0.19***	-0.35***	-0.47***
	(2.96)	(0.95)	(-3.07)	(-3.31)	(-4.66)	(-5.15)
4	0.09**	0.04*	-0.13***	-0.2***	-0.36***	-0.45***
	(2.41)	(1.67)	(-2.75)	(-3.88)	(-4.12)	(-4.36)
5	0.09**	0.04	-0.11**	-0.19***	-0.4***	-0.49***
	(2.24)	(1.34)	(-2.53)	(-3.69)	(-3.99)	(-4.05)
6	0.07	0.04	-0.09**	-0.13***	-0.24***	-0.31***
	(1.64)	(1.25)	(-2.4)	(-2.67)	(-3.03)	(-2.96)
7	0.06	0.04	-0.08**	-0.13***	-0.25***	-0.32***
	(1.48)	(1.58)	(-2.3)	(-2.68)	(-3.06)	(-2.89)
8	0.07*	0.03	-0.05	-0.09*	-0.29***	-0.36***
	(1.86)	(1.04)	(-1.29)	(-1.76)	(-2.82)	(-2.85)
9	0.08*	0.01	-0.01	-0.1**	-0.26***	-0.34***
	(1.92)	(0.31)	(-0.27)	(-1.98)	(-2.7)	(-2.86)
10	0.07	0.01	0.01	-0.07	-0.28**	-0.35**
	(1.59)	(0.38)	(0.18)	(-1.23)	(-2.44)	(-2.57)
10-0	-0.04**	-0.07**	0.25***	0.32***	0.4***	0.44***
	(-2.36)	(-2.24)	(4.91)	(4.4)	(4.92)	(5.09)

Table C.9-continued

Horizon (years)	Panel E. EG					
	0	1	2	3	4	HL
0	-0.18***	-0.03	0.28***	0.29***	0.64***	0.82***
	(-4.88)	(-0.9)	(5.54)	(3.8)	(5.92)	(6.52)
1	-0.16***	-0.03	0.26***	0.27***	0.4***	0.56***
	(-4.2)	(-0.82)	(4.92)	(3.71)	(3.06)	(3.63)
2	-0.17***	0.02	0.21***	0.19***	0.29***	0.46***
	(-4.29)	(0.6)	(4.08)	(2.87)	(2.59)	(3.3)
3	-0.17***	0.05	0.2***	0.14**	0.25**	0.42***
	(-4.19)	(1.39)	(3.74)	(2.04)	(2.18)	(3.0)
4	-0.16***	0.07*	0.17***	0.15**	0.32***	0.48***
	(-3.95)	(1.85)	(3.28)	(2.35)	(2.85)	(3.49)
5	-0.15***	0.08*	0.11**	0.19***	0.35***	0.5***
	(-4.11)	(1.76)	(2.2)	(3.18)	(3.25)	(3.81)
6	-0.14***	0.11***	0.12***	0.12**	0.29***	0.44***
	(-3.65)	(2.65)	(2.63)	(2.02)	(2.66)	(3.22)
7	-0.14***	0.12***	0.11***	0.15***	0.23*	0.37***
	(-3.68)	(3.41)	(2.91)	(2.64)	(1.95)	(2.6)
8	-0.13***	0.15***	0.06	0.13**	0.23*	0.36**
	(-3.35)	(3.28)	(1.46)	(2.3)	(1.85)	(2.39)
9	-0.13***	0.15***	0.08**	0.11*	0.18	0.31**
	(-3.21)	(3.72)	(2.03)	(1.81)	(1.44)	(2.09)
10	-0.12***	0.19***	0.04	0.08	0.17	0.29**
	(-2.84)	(4.56)	(0.8)	(1.35)	(1.57)	(2.18)
10-0	0.06***	0.21***	-0.25***	-0.21***	-0.46***	-0.53***
	(2.9)	(5.37)	(-4.01)	(-2.84)	(-3.74)	(-4.07)

Table C.9-continued