Does Peer-Reviewed Research Help Predict Stock Returns?

Andrew Y. Chen[1], Alejandro Lopez-Lira[2], and Tom Zimmermann[3]

[1]Federal Reserve Board
[2]University of Florida
[3]University of Cologne and Centre for Financial Research

June 2024

Abstract

Mining 29,000 accounting ratios for t-statistics over 2.0 leads to cross-sectional return predictability similar to the peer review process. For both methods, about 50% of predictability remains after the original sample periods. Predictors supported by peer-reviewed risk explanations or equilibrium models underperform other predictors post-sample, suggesting peer review systematically mislabels mispricing as risk, though only 20% of predictors are labelled as risk. Data mining generates other features of peer review including the rise in returns as original sample periods end and the speed of post-sample decay. It also uncovers themes like investment, issuance, and accruals—decades before they are published.

---

1 Introduction

Suppose a Ph.D. student tells you he found a predictor with a long-short return of 100 bps per month in a historical sample. You ask him, "where does this predictor come from?" How would your view about the post-sample return change if the predictor is:

1. Based on an idea that is publishable in a top finance journal (e.g. Journal of Finance)

2. Found by mining tens of thousands of accounting ratios for t-stats greater than 2.0?

In other words, how much does peer-reviewed research help predict cross-sectional returns compared to naive data mining?

To answer this question, we construct the empirical counterpart to the scenario. We match 200 published predictors (from Chen and Zimmermann (2022)) to data-mined benchmarks. The data-mined benchmarks come from searching 29,000 accounting ratios for t-stats greater than 2.0 in the published predictors' original sample periods. The accounting ratios are naive: they are simply ratios or scaled first differences using 240 Compustat accounting variables (+ CRSP market equity). The only restriction on these ratios is that we avoid dividing by variables that are typically zero or negative. We form long-short portfolios for each predictor and re-scale so that the mean original-sample return is 100 bps per month. Finally, we compare post-sample returns.

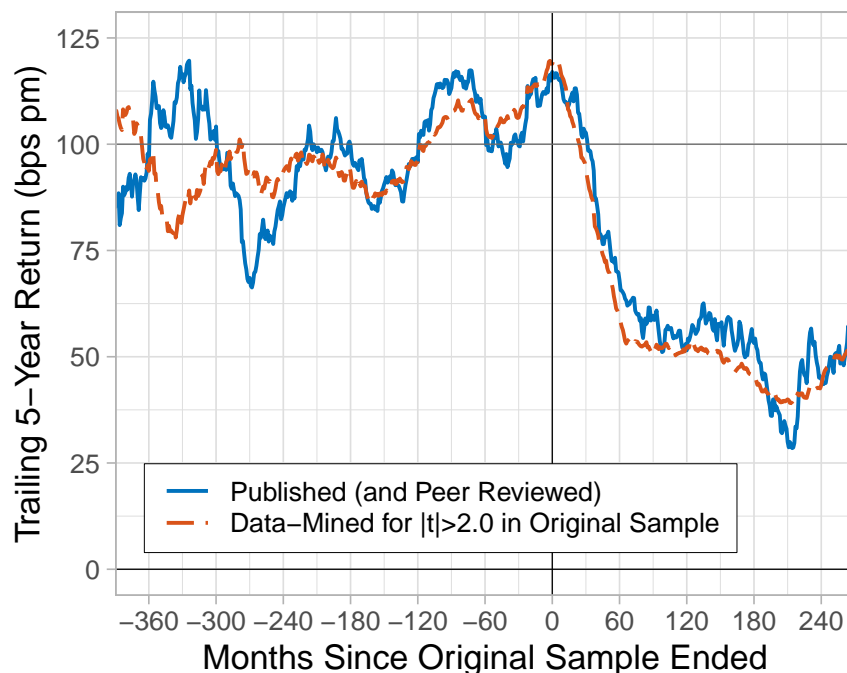Figure 1: Does Peer-Reviewed Research Help Predict Returns?

1

Figure 1 illustrates the result. It plots returns of published and data-mined strategies in event time, where the event is the month that the original sample ended. Returns are averaged across strategies within each event-time month and then the trailing 5-year return is taken for ease of reading. As shown in the seminal McLean and Pontiff (2016) meta-study, published returns (solid line) decay post-sample, but they remain far above zero, averaging 53% of their original sample means. Data-mined returns (dashed line) decay a bit more, with post-sample means that are 51% of their original sample means. So peer-reviewed research seems to help predict returns compared to data mining, but the improvement is modest. The publishable predictor in our hypothetical scenario outperforms by only 2 bps per month.

In fact, Figure 1 shows that it is hard to reject the null that the academic discovery process is, itself, data mining. Data mined returns match not only the post-sample decay of published returns: they also match the rise in trailing 5-year returns as the original samples end, the decline in returns in the first 60 months post-sample, the flattening of returns in months 60 to 120, and the dip in returns around month 210.

Data mining even uncovers the same themes as academic research. The accounting ratios that generate the most statistically-significant predictability in the 1963-1980 sample are related to investment (Titman, Wei, and Xie (2004)), debt issuance (Spiess and Affleck-Graves (1999)), share issuance (Pontiff and Woodgate (2008)), accruals (Sloan (1996), inventory growth (Thomas and Zhang (2002)), and earnings surprise (Foster, Olsen, and Shevlin (1984)). These themes are consistently found in more recent samples too. Thus, data mining not only uncovers a similar event-time returns, it also uncovers the same ideas as peer-reviewed research. Shockingly, data mining can uncover these themes *decades* before they are published.

Perhaps risk-based research can identify predictors that outperform. As described in Cochrane's (2009) influential textbook, "the best hope for finding pricing factors that are robust out of sample and across different markets, is to try to understand the fundamental macroeconomic sources of risk." Many of the papers in the Chen and Zimmermann (2022) dataset do not follow this advice, and motivate their predictors using informal arguments about mispricing. Some even lack a clear explanation, and base their conclusions on the strength of their empirical results. For example, Banz (1981) ends with "the size effect exists but it is not at all clear why it exists."

To address this possibility, we assign predictors to "risk," "mispricing," or "agnostic" categories based on the explanation for predictability in the original papers. We then compare the post-sample returns of each group.

The main result remains: risk-based research does not lead to higher post-sample

returns compared to data mining. If anything, risk-based predictors underperform their data-mined benchmarks. We find similar results when categorizing predictors based on the support of a mathematical equilibrium model. While there are relatively few predictors supported by formal models, the ones that exist imply that the relationship between modeling rigor and post-sample returns is *negative*.

An important caveat is that our results characterize cross-sectional predictability research as it was practiced from 1973 to 2016, the publication years covered by the Chen and Zimmermann (2022) dataset. Research evolves over time, and since 2016, a growing number of academics have embraced machine learning and other big data methods (Moritz and Zimmermann (2016); Yan and Zheng (2017); Gu, Kelly, and Xiu (2020)).[1] Outside of finance, fields like protein folding and language modeling have been recently revolutionized by atheoretical searches through vast amounts of data (Jumper et al. (2021); Zhao et al. (2023)).

Figure 1 is a stark illustration of the promise of big data. Simply searching accounting variables for large t-statistics generates substantial out-of-sample returns. And while data-mined results say little by themselves about the underlying economics, they can provide the empirical foundation for the next generation of economic ideas (Chen and Dim (2023)).

As a secondary result, we document a striking consensus about the origins of cross-sectional predictability, according to peer review. Among the 199 published predictors we examine, only 18% are attributed to risk by peer review. 59% are attributed to mispricing, and 23% have uncertain origins.

This consensus is a positive sign regarding the scientific process in finance. The fact that risk-based predictors consistently decay post-sample implies that peer review either mislabels mispricing as risk or identifies unstable risk factors that weaken over time. Fortunately, these errors are uncommon, and represent a relatively small "false discovery rate."

A more negative view of the scientific process comes from the fact that recent reviews of cross-sectional predictability are agnostic about risk vs mispricing (Bali, Engle, and Murray (2016); Zaffaroni and Zhou (2022)). Given the strong consensus found from reading the individual papers, this agnosticism suggests that the battle between risk-based and behavioral finance has led to an unwillingness to engage in debate. This

---

[1]An incomplete list of additional big data papers includes Green, Hand, and Zhang (2017); DeMiguel et al. (2020); Freyberger, Neuhierl, and Weber (2020); Kozak, Nagel, and Santosh (2020); Han et al. (2022); Chen and Velikov (2022); Bessembinder, Burt, and Hrdlicka (2023); Lopez-Lira and Roussanov (2020); Jensen, Kelly, et al. (2022); and Chen and McCoy (2024).

unwillingness raises questions about whether the field of asset pricing is self-correcting (Ankel-Peters, Fiala, and Neubauer (2023)) and whether peer review has the power to reject false paradigms (Akerlof and Michaillat (2018)).

1.1 Related Literature

To our knowledge, our paper is the first to test the widely-held belief that economic theory improves out-of-sample robustness relative to data mining. In previous research, this belief is either assumed to be true (Harvey, Liu, and Zhu (2016); Harvey (2017); Fama and French (2018)) or expressed as a "best hope" (Cochrane (2009)). Our tests provide evidence inconsistent with this belief—if theory is practiced the way it was in the papers covered by the Chen and Zimmermann (2022) dataset. We also provide a meta-theory for understanding why theory may fail to improve robustness.

Earlier papers on data mining studied statistical theory (Lo and MacKinlay (1990), see also Chen (2021)) or data mining for time-series predictability (Sullivan, Timmermann, and White (1999); Sullivan, Timmermann, and White (2001)). Our paper fits in with the more recent literature following Yan and Zheng (2017), which mines for cross-sectional predictability (Chordia, Goyal, and Saretto (2020); Harvey and Liu (2020); Goto and Yamada (2022); Zhu (2023); Chen (2024)). Relative to these papers, ours is unique in showing how closely data mining resembles peer-reviewed research. We are also unique in focusing on out-of-sample tests, which are well-understood and have straightforward interpretations. The aforementioned papers focus on multiple testing methods, which can be easily misinterpreted (Chen and Zimmermann (2023)). Following up on our paper, Chen and Dim (2023) show how to use empirical Bayes to mine more rigorously.

Our paper provides a new angle on the risk vs mispricing debate in the cross-section of stock returns (Fama (1970); Shiller (2003); Cochrane (2017); Barberis (2018); etc). Since Fama (1970), it has been recognized that standard empirical tests can only reject special cases of the broader class of risk theories (the "joint hypothesis problem"). Our methods attack this problem by building on the efforts of the asset pricing community. This community is, in a way, an organic computer designed to search the entire class of risk theories. Based on our tests, this search has uncovered little robust cross-sectional risk during the years covered by the Chen and Zimmermann (2022) dataset. Our use of peer-reviewed judgment is also distinct from contemporaneous papers that aim to classify anomalies in terms of risk vs mispricing, which focus on option market volume (Böll et al. (2024)), analyst forecasts revisions (Frey (2023)), stochastic dominance (Holcblat, Lioui, and Weber (2022)), or optimized factor models (Bali, Beckmeyer, and Wiedemann (2023)).

Other papers that compare peer-reviewed to alternative predictors include Soebhag, Van Vliet, and Verwijmeren (2024); Hasler (2023); and Walter, Weber, and Weiss (2023). The alternative predictors in these papers are perturbations of the peer-reviewed predictors, while our alternatives are generated by sheer data mining.

2 Data-Mined Predictability

We describe our data mining procedure and the predictability it uncovers.

2.1 A Simple Data Mining Procedure

We begin with 241 Compustat annual accounting variables examined by Yan and Zheng (2017). Yan and Zheng select these variables to (1) ensure non-missing values in at least 20 years and (2) that the average number of firms with non-missing values is at least 1,000 per year. We add CRSP market equity, leading to 242 "ingredient" variables.

We then generate 29,315 accounting ratios (signals) using two functional forms: simple ratios (X/Y) and first differences scaled by a lagged denominator ($\Delta X/\text{lag}(Y)$). The numerator can use any of the 242 ingredients. The denominator is restricted to the 65 ingredients that are positive for at least 25% of firms in 1963 with matched CRSP data. This restriction avoids normalizing by zero or negative numbers. This procedure leads to $\approx 242 \times 65 \times 2 = 31,460$ ratios, but we drop 2,145 ratios that are redundant in "unsigned" portfolio sorts.[2]

We lag each signal by six months relative to the fiscal year ends, and then form long-short decile strategies by sorting stocks on the lagged signals in each June. Delisting returns and other data handling methods follow Chen and Zimmermann (2022). For further details, please see `https://github.com/chenandrewy/flex-mining`.

In our view, this process is the simplest reasonable data mining procedure. A reasonable data mining procedure should include both ratios and first differences. Scaling first differences by a lagged variable nests percentage changes, which likely should also be included in a reasonable data mining process.

This procedure is inspired by Yan and Zheng (2017), who create 18,000 accounting ratios using transformations inspired, in part, by the asset pricing literature. Choosing transformations based on the literature could potentially lead to look-ahead bias, which

---

[2]For the $65 \times 65 = 4,225$ ratios where the numerator is also a valid denominator, there are only 65 choose 2 = 2,080 ratios that are in a sense distinct.

our procedure avoids. However, previous versions of this paper used Yan and Zheng's data and found similar results.

2.2 Out-of-Sample Returns from Data Mining

Our simple data mining procedure generates notable out-of-sample returns, as seen in Table 1. For each June starting in 1992, we sort the data-mined strategies into five bins based on their mean returns over the past 30 years (in-sample) and compute the mean return over the next year within each bin (out-of-sample). We then average these statistics across each June.

Table 1: Out-of-Sample Returns from Mining Accounting Data

We sort 29,000 data-mined long-short strategies each June (1992-2021) into 5 bins based on past 30-year mean returns (in-sample) and compute the mean return over the next year within each bin (out-of-sample). Statistics are calculated by strategy, then averaged within bins, then averaged across sorting years. Decay is the percentage decrease in mean return out-of-sample relative to in-sample. We omit decay for bin 4 because the mean return in-sample is negligible. Data-mined returns are large and comparable to published returns, both in- and out-of-sample.

| In-Sample Bin | Equal-Weighted Long-Short Deciles | | | | Value-Weighted Long-Short Deciles | | | |
| | Past 30 Years (IS) | | Next Year (OOS) | | Past 30 Years (IS) | | Next Year (OOS) | |
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

Using equal-weighted strategies, the in-sample returns of the first bin are on average -59 bps per month, with an average t-stat of -4. These statistics are similar to those of the typical published predictor (Chen and Zimmermann (2022)). Out-of-sample, the first bin returns -49 bps per month, implying a decay of only 17%, once again resembling published predictability (McLean and Pontiff (2016)). Since investors can flip the long and short legs of these strategies, these statistics imply substantial out-of-sample returns. Similar predictability is seen in bin 5, which decays by 27%.

Out-of-sample predictability is also seen in value-weighted strategies, though the magnitudes are weaker. Still, the out-of-sample returns monotonically increase in the

in-sample return, indicating the presence of true predictability. Moreover, the roughly 60% decay is far from 100%, and is in the ballpark of the post-sample decay for published predictors (McLean and Pontiff (2016)). Similarly, out-of-sample predictability is much weaker post-2004, though it still exists (see Appendix Table A.1).

Since there are 29,000 data-mined signals, Table 1 implies thousands of strategies with notable out-of-sample predictability. But are these strategies distinct? To address this question, we describe the covariance structure of data-mined predictors in Table 2. The table examines strategies that have t-stats greater than 2.0 in at least 10% of the 30-year in-sample periods from Table 1.

Table 2: Correlation Structure of Data-Mined Predictors

We characterize the covariance structure of data-mined predictor returns over the 1994-2020 sample. Data-mined predictors are represented by strategies with t-statistics greater than 2.0 in at least 10% of the in-sample periods from Table 1. Panel (a) reports percentiles of Pearson correlation coefficients computed over pairwise-complete return observations. Panel (b) uses principal component analysis on strategies with no missing returns in the 1994-2020 sample. Data-mining uncovers many distinct predictors.

Panel (a): Pairwise correlations									
Percentile	1	5	10	25	50	75	90	95	99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)											
Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

Table 2 shows that the predictors are to a significant extent distinct. More than 80% of pairwise correlations are below 0.25 in absolute value (Panel (a)). Many dozens of principal components are required to span 80% of total variance (Panel (b))—though there is a non-trivial factor structure. Thus, data mining not only uncovers notable out-of-sample performance, but also generates a very large number of distinct strategies. A similar covariance structure is seen in published predictors (Chen and Zimmermann (2022); Bessembinder, Burt, and Hrdlicka (2023)).

2.3 Data-Mined Predictability Themes

To study themes, we manually categorize the accounting ratio numerators that produce the largest t-stats in the 1963-1980 sample. 1980 is the year B/M is published (Stattman (1980)) and only five other predictors from the Chen and Zimmermann (2022) dataset have been published by this time (Beta, Price, Earnings/Price, and Dividend Yield Short-Term). Thus, the themes found in this analysis are largely unspanned by the literature as of 1980. Nevertheless, we find similar themes using samples ending in 1990, 2000, and 2010 (Appendix Tables A.2-A.4).

Table 3 reports the 20 numerator and stock weight (equal- or value-) combinations that produce the largest mean t-stats in the 1963-1980 sample, where the mean is taken across the 65 possible denominators. We then manually assign the numerators to themes.

All of the top 20 numerators fit into themes from the cross-sectional literature. These themes include investment (Titman, Wei, and Xie (2004)), debt issuance (Spiess and Affleck-Graves (1999)), share issuance (Pontiff and Woodgate (2008)), accruals (Sloan (1996)), inventory growth (Thomas and Zhang (2002)), and earnings surprise (Foster, Olsen, and Shevlin (1984)). For all of these themes, the sign of predictability obtained from data mining is the same as the sign from the literature (e.g. short stocks with high investment).

Thus, data mining works, in part, by uncovering the same ideas found by peer review. One may have thought that a deep understanding of financial economics is required to uncover these themes. But it turns out that mining accounting data for large t-stats is sufficient. In fact, data mining could have uncovered these themes *decades* before they were published.

One might also think that data mining would uncover spurious themes, given the warnings about data mining going back to Jensen and Benington (1970). However, every single one of the numerators in Table 3 produces returns that persist out-of-sample. In fact, during the 1981-2004 out-of-sample period, the return decay is on average zero.

Returns decay notably post-2004, and on average are only 40% as large as they were pre-1981. This decay is also seen in published predictors (Chen and Velikov (2022)), and has been attributed to the rise of the internet (Chordia, Subrahmanyam, and Tong (2014)) as well as learning from academic publications (McLean and Pontiff (2016)).

Taken together, these results hint at our main finding. Data-mined predictability is remarkably similar to that of peer-reviewed research. This resemblance is seen in performance both in- and out-of-sample (Table 1), correlation structure (Table 2), and even themes (Table 3). These results also illustrate the power of data-driven methods. The simplest reasonable data mining method generates substantial out-of-sample returns,

Table 3: Themes from Mining Accounting Ratios in 1980

Table reports the 20 accounting ratio numerator and stock weight (equal- or value-) combinations with the largest mean t-stats using returns in the years 1963-1980 (IS). 'ew' is equal-weight, 'vw' is value-weight. We manually group numerators into themes from the literature. Strategies are signed to have positive mean returns IS. 'Pct Short' is the share of strategies that short stocks with high ratios. 't-stat' and 'Mean Return' are averages across the 65 possible denominators. 'Mean Return' is in bps per month. 'Mean return OOS/IS' is the mean in either 1981-2004 or 2005-2022 (OOS), divided by the mean IS. Data mining can uncover themes from the literature like investment, external financing, and accruals, decades before they are published. For all themes, predictability persists out-of-sample.

| Numerator (Stock Weight) | 1963-1980 (IS) | | | 1981-2004 | 2005-2022 |
	Pct Short	t-stat	Mean Return	Mean Return OOS / IS	
Investment / Investment Growth (Titman, Wei, Xie 2004; Cooper, Gulen, Schill 2008)					
ΔAssets (ew)	100	4.0	0.86	1.05	0.34
ΔIntangible assets (ew)	100	4.0	0.52	1.04	0.30
ΔPPE net (ew)	98	4.0	0.78	1.09	0.22
ΔPPE gross (ew)	98	3.8	0.76	1.00	0.18
ΔInvested capital (ew)	100	3.5	0.73	1.35	0.39
ΔCapital expenditure (ew)	100	3.2	0.43	1.54	0.51
External Financing (Spiess and Affleck-Graves 1999; Pontiff and Woodgate 2008)					
ΔCommon stock (ew)	100	5.1	0.80	0.67	0.30
ΔLiabilities (ew)	100	4.7	0.80	0.79	0.30
ΔCapital surplus (ew)	100	4.1	0.60	1.21	0.97
ΔLong-term debt (ew)	100	3.6	0.47	1.43	0.36
ΔCapital surplus (vw)	98	3.0	0.54	0.93	0.60
Accruals / Inventory Growth (Sloan 1996; Thomas and Zhang 2002; Belo and Lin 2012)					
ΔInventories (ew)	100	4.2	0.66	1.22	0.28
ΔNotes payable st (ew)	100	3.8	0.43	0.59	0.34
ΔReceivables (ew)	100	3.7	0.67	0.59	0.41
ΔDebt in current liab (ew)	100	3.7	0.43	0.73	0.35
ΔCurrent liabilities (ew)	100	3.7	0.51	1.33	0.21
Earnings Surprise (Foster, Olsen, Shevlin 1984; Chan, Jegadeesh, Lakonishok 1996)					
ΔCost of goods sold (ew)	100	3.7	0.60	0.87	0.32
ΔOperating expenses (ew)	98	3.5	0.58	0.96	0.45
ΔSG&A (ew)	100	3.3	0.62	1.04	0.36
ΔInterest expense (ew)	98	3.3	0.48	1.35	0.76

from a variety of distinct sources.

3   Research vs Data Mining

We compare the post-sample returns of peer-reviewed research to data mining.

3.1   Peer-Reviewed Predictors Data

Peer-reviewed predictors come from the August 2023 release of the Chen and Zimmermann (2022) (CZ) dataset. This dataset is built from 212 firm-level variables that were shown to predict returns cross-sectionally in academic journals. It covers the vast majority of firm-level predictors that can be created from widely-available data and were published before 2016. The CZ data is a uniquely accurate representation of the literature: unlike other large-scale replications, CZ show that their t-stats are generally a good match for the t-stats in the original papers.

We drop five predictors that produce mean long-short original-sample returns of less than 15 bps per month in CZ's replications. These predictors are rather distant from the original papers.[3] Since these predictors are rare, including them has little effect on our results. We also drop 8 predictors that have less than 9 years of post-sample returns. The median post-sample length for the remaining predictors is about 20 years.

3.2   Post-Sample Returns: Research vs Data Mining

We can now answer the question posed on page 1. How much does peer-reviewed research help predict cross-sectional returns compared to data mining?

To answer this question, we construct data-mined benchmarks. For each published predictor, we search the 29,000 accounting ratios for long-short strategy absolute t-stats > 2.0, using the same sample period and stock weighting as the original paper. This generates roughly 6,000 data-mined benchmarks for each published predictor. We then compare the returns of the published strategies with their average benchmark.

Figure 1 shows the result. It plots the returns of published predictors and data-mined benchmarks in event time, where the event is the end of the published predictors' original

---

[3]For example, CZ equal-weight the Frazzini and Pedersen (2014) betting against beta portfolios instead weighting by betas. CZ use CRSP age rather than the NYSE archive data used by Barry and Brown (1984). CZ also find very small returns in simple long-short strategies for select variables shown by Haugen and Baker (1996), Abarbanell and Bushee (1998), Soliman (2008) to predict returns in multivariate settings.

sample periods. All strategies are normalized to have 100 bps mean return in the original samples for ease of interpretation. Returns are then averaged across predictors by group (published or data-mined) and then the trailing 5-year mean is taken to smooth out noise. We use event-month returns for data-mined benchmarks only if the corresponding published return is observed, to ensure that the returns are comparable. Section 5.1 shows the 100 bps normalization has little effect on our results. We also find similar results if we limit the published predictors to those that focus on accounting data (Appendix Figure A.2).

Post sample, peer-reviewed (solid line) and data-mined (long-dash) predictors perform similarly. In fact, research and data mining lead to eerily similar event-time returns, with the data-mined returns resembling a Kalman-filtered version of the research returns. In this sense, it is difficult to reject the null that the research process is built off of data mining—at least for the research covered in the Chen and Zimmermann (2022) meta-study.

3.3 Even More Naive Data Mining Methods

Our data mining process (Section 2.1) just searches accounting ratios for t-stats > 2.0. But one can think of even more naive methods. How naive can one be and still generate research-like out-of-sample returns?
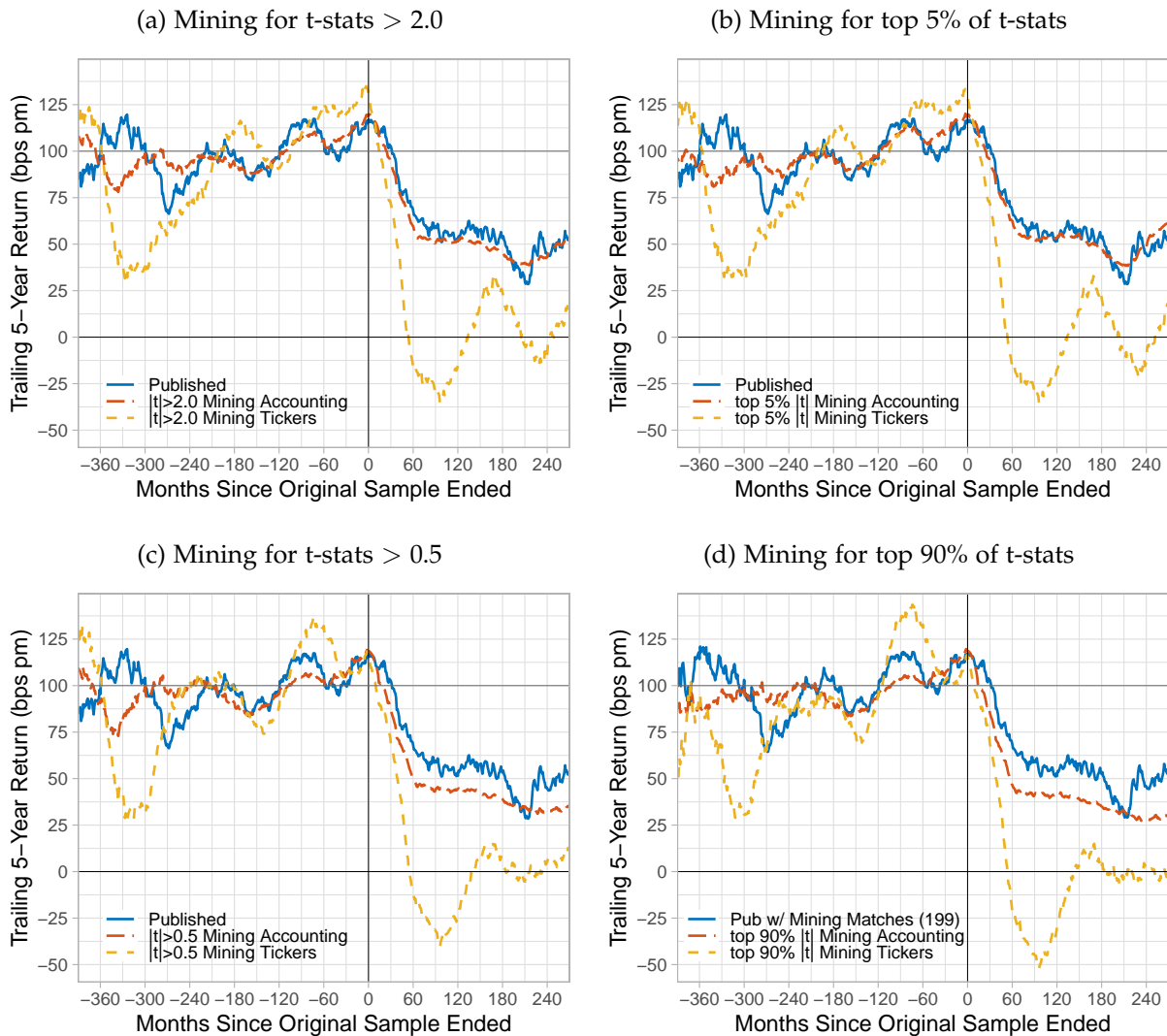
To answer this question, we examine an alternative data mining method proposed in Harvey (2017). Harvey asks his research assistant to "form portfolios based on the first, second, and third letters of the ticker symbol," leading to 3,160 long-short portfolios. We interpret his instructions as follows: Generate 26 portfolios by going long all stocks with a first ticker letter of "A," "B," "C," ..., and "Z." Generate 26 portfolios by doing the same for the second ticker letter, and add a 27th portfolio for tickers that have no second ticker letter. Apply the same to the third ticker. This process results in $26 + 27 + 27 = 80$ long portfolios. Finally, form $\binom{80}{2} = 3,160$ long-short portfolios by selecting all distinct pairs of the 80 long portfolios.

Figure 2 compares this ticker mining procedure to our baseline mining of accounting ratios. Panel (a) applies the same selection procedure as in Figure 1: we screen for t-stats > 2.0 in the original sample periods. Post-sample, the mean returns from mining tickers (short-dash line) is approximately zero. Thus, mining tickers is too naive.

Panel (b) applies an alternative selection procedure: we screen for the top 5% of t-stats in the original sample periods. Mining accounting data still leads to very similar returns as research—in fact, the returns even closer to research than those of the t-stat > 2.0

Figure 2: Even More Naive Data Mining Methods

We compare published predictors (solid) to benchmarks made from data mining accounting ratios (long-dash) or tickers (short-dash). Benchmarks screen for a minimum t-stat (Panels (a) and (c)) or for the top X% of data-mined t-stats (Panels (b) and (d)) in the published papers' original sample periods. The plot shows the long-short returns in event time, where the event is the end of the original sample periods. Each predictor is normalized so that its mean original-sample return is 100 bps per month. Returns are averaged across predictors (by group) within each event month and then the trailing 5-year average is taken for readability. Extremely naive data mining generates post-sample returns in the ballpark of research, though mining tickers is too naive.

(a) Mining for t-stats > 2.0   (b) Mining for top 5% of t-stats

(c) Mining for t-stats > 0.5   (d) Mining for top 90% of t-stats

screen. This result suggests that research does not just screen for statistical significance—it instead aims for the strongest signals available in the data. Mining tickers for the top 5% of t-stats still leads to post-sample returns of around zero.

The bottom panels examine much more lenient statistical screens. Panel (c) screens for t-stats > 0.5 and Panel (d) screens for the top 90% of t-stats (enter all strategies except for the worst 10%). Since we sign strategies to have positive original sample returns, these screens effectively take on only sign information (flip the long and short legs if the strategy had negative mean returns in the past). The result is underperformance relative to research, but the difference is moderate. Mining accounting data still leads to post-sample returns that are about 75% as large as those from research.

Figure 2 illustrates two lessons about data mining. The first is that the data being mined is important. Some data, like accounting ratios, are full of informative signals. Mining such data generates out-of-sample returns, even if the mining process uses only the sign of past mean returns. Other data, like tickers, are entirely uninformative. Mining these data only uncovers noise, and returns that vanish out-of-sample.

The second lesson is that more data mining does not necessarily mean worse out-of-sample performance. The accounting dataset is almost 10 times as large as the ticker dataset, yet it produces much stronger post-sample returns. In fact, we show that the amount of mining is actually irrelevant in our model of the "Cochrane Conjecture" (Section 4.4).

In summary, naively mining accounting data leads to post sample returns that are remarkably similar to those from the peer review process. While this result is negative for peer review, it illustrates the promise of big data methods. These methods can identify true predictability, even when applied in the most naive ways.

4   Does Risk-Based Theory Help?

This section examines whether research that focuses on risk-based, equilibrium forces can find more stable returns, and thus outperform data mining.

4.1   Risk or Mispricing? According to Peer Review

To study risk-based research, we categorize predictors as risk-based, mispricing-based, or agnostic using the texts in the original papers. We manually read each paper, identify a passage of text that summarizes the main argument, and then categorize the passage. The passages are typically taken from either the abstract, introduction, or conclusion. Each paper was reviewed by two of the authors.

Table 4 provides a representative passage for each category. The risk and mispricing passages are straightforward: risk passages discuss risk, equilibrium, or market efficiency,

while mispricing passages discuss mispricing or investor errors. Agnostic passages are slightly more difficult. Agnostic predictors are easy to classify when the papers claim agnosticism or provide arguments for both risk and mispricing. But in some cases, agnostic papers avoid discussing the explanation for predictability at all, and instead focus on the empirics. For example, Boudoukh et al. (2007) provide extensive evidence on the predictive power of payout yield and the importance of measuring repurchases as well as dividends, but do not explicitly argue for a risk or mispricing explanation.

Table 4: Risk or Mispricing? According to Peer Review

We classify predictors into "risk," "mispricing," or "agnostic" by identifying passages that summarize the main argument in the corresponding papers and then classifying the passage. All passages and their classifications are found at `https://github.com/chenandrewy/flex-mining/blob/main/DataInput/SignalsTheoryChecked.csv`. "JF, JFE, RFS" includes only predictors published in the Journal of Finance, Journal of Financial Economics, or Review of Financial Studies. Peer review attributes only about 20% of predictors to risk.

| Category | Num Predictors | | Example Predictor | Example Passage |
	Any Journal	JF, JFE, RFS		
Risk	36	33	Real estate holdings (Tuzel 2010)	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.
Mispricing	117	65	Share repurchases (Ikenberry, Lakonishok, Vermaelen 1995)	The market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements
Agnostic	46	25	Size (Banz 2981)	To summarize, the size effect exists but it is not at all clear why it exists
Total	199	123		

In a handful of cases, the text argues for liquidity explanations. We categorize these predictors as mispricing if the argument focuses on stock-specific measures of liquidity (Amihud (2002)) and risk if the argument focuses on a market-wide component (Pástor and Stambaugh (2003)). This method gives the risk category the best chance at finding post-sample returns, since idiosyncratic liquidity has improved over time. Nevertheless, this issue affects only seven predictors, and has little impact on our main results.

This meta-analysis finds a remarkable consensus about the origins of cross-sectional

predictability. Table 4 shows that only 18% of cross-sectional predictors are judged by the peer review process to be due to risk. In contrast, 59% of predictors are due to mispricing. The remaining 23% of predictors are agnostic. Top finance journals seem to favor risk-based explanations but still attribute only 26% of predictors to risk. A detailed breakdown by journal is in Appendix Table A.5.

The strong consensus in Table 4 contrasts with recent reviews on empirical cross-sectional asset pricing (e.g. Bali, Engle, and Murray (2016) and Zaffaroni and Zhou (2022)). These reviews provide a largely agnostic description of the origins of predictability, suggesting that peer review has come to a divided view, or that this topic has become too contentious for open discussion. Our results show that the literature favors mispricing, and that a minority of predictors are due to risk, as judged by the peer review process.

4.2 Post-Sample Returns of Risk vs Mispricing

Figure 3 shows the post-sample returns of risk-based, mispricing-based, and agnostic predictors. As in the previous figures, we plot trailing 5-year long-short returns in event time, and normalize each strategy to have 100 bps mean return in the original samples.

Risk-based predictors actually decay *more* than other predictors. The figure gives the illusion of outperformance in the first few years post-sample, but the trailing 5-year return is not fully post-sample until month 60 in the plot. This result is robust to adjusting for CAPM exposure (Appendix Figure A.3), though we focus on raw returns because the CAPM typically holds in risk-based models of cross-sectional predictability (e.g. Zhang (2005); Tuzel (2010)). The underperformance in Figure 3 comes from just 36 risk-based predictors, which raises questions about statistical significance.

Table 5 examines statistical significance in a regression framework (following McLean and Pontiff (2016)). Specification (1) regresses monthly long-short returns on a post-sample indicator and its interaction with an indicator for risk-based predictors. Returns are normalized to be 100 bps per month in-sample, so the post-sample coefficient implies that returns decay by 42 percent overall (across all types of predictors).[4] The interaction coefficient implies that risk-based predictors have an additional decay of 29 percentage points, for a total decay of 71 percent. The additional decay of risk predictors is only marginally significant, however.

Nevertheless, there is plenty of data to show that risk-based explanations fail to

---

[4]This decay implies 58% of returns remains post-sample if each predictor-month is weighed equally. This is a bit higher than the 53% found on page 1, which weighs each month equally, and thus focuses more on returns further from the original sample.

Figure 3: Post-Sample Returns by Peer-Reviewed Explanation

The plot shows the mean long-short returns of published predictors in event time, where the event is the end of the original sample periods. Each predictor is normalized so that its mean in-sample return is 100 bps per month. Predictors are grouped by theory category based on the arguments in the original papers (Table 4). We average returns across predictors within each month and then take the trailing 5-year average for readability. For all categories of theory, predictability decays by roughly 50% post-sample. Risk-based predictors decay more than other predictors.
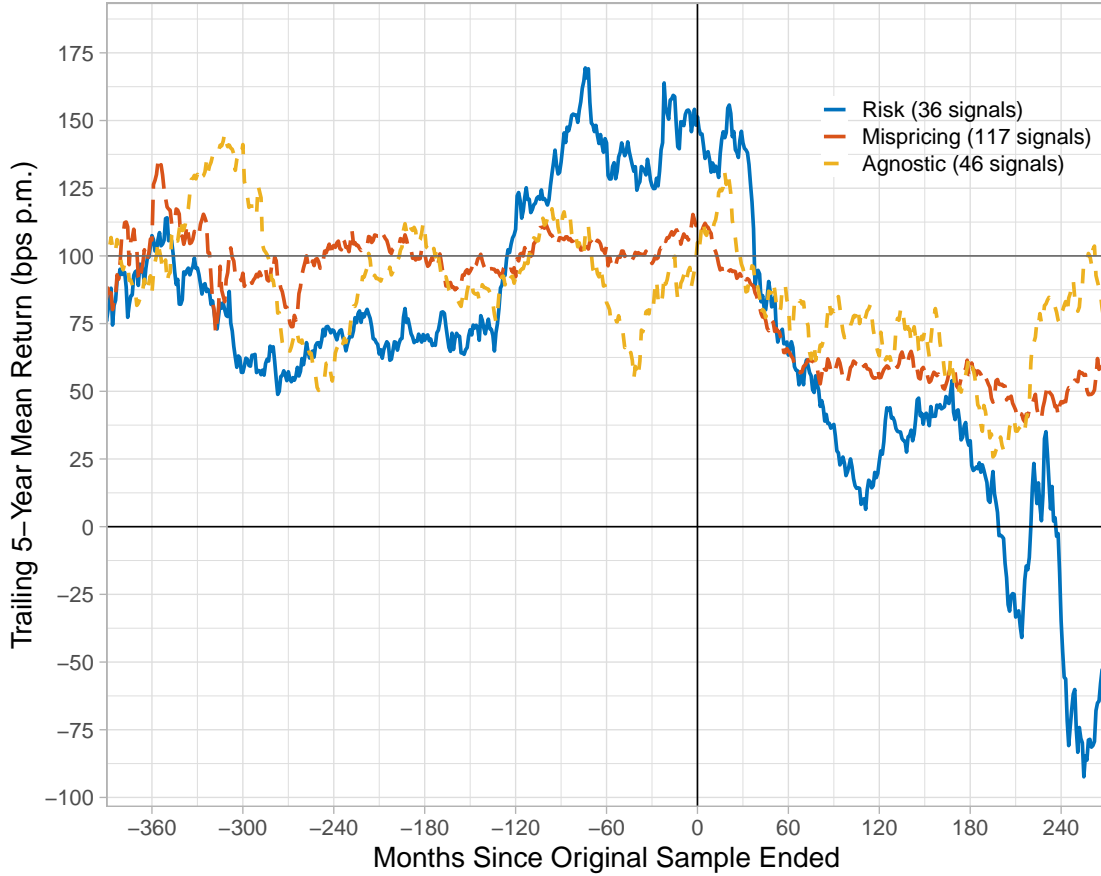
prevent post-sample decay. This result is shown in the row "Null: Risk No Decay," which tests the hypothesis that the sum of the Post-Sample and Post-Sample × Risk coefficients is non-negative. The test rejects this hypothesis at the 0.1% level.

Specifications (2)-(4) show robustness. Specification (2) adds a post-publication indicator, specification (3) adds an indicator for mispricing explanations, and specification (4) adds both. All three alternative specifications arrive at risk-based predictors decaying by an additional 30 to 40 percentage points. Specification (4) implies that post-publication, being risk-based implies an additional 20 + 20 = 40 percentage points of decay, for a total decay of 24 + 15 + 40 = 79%.

Table 5: Regression Estimates of Risk vs Mispricing Effects on Predictability Decay

We regress monthly long-short returns on indicator variables to quantify the effects of peer-reviewed risk vs mispricing explanations on predictability decay. Each strategy is normalized to have 100 bps per month returns in the original sample. "Post-Sample" is 1 if the month occurs after the predictor's sample ends and is zero otherwise. "Post-Pub" is defined similarly. "Risk" is 1 if peer review argues for a risk-based explanation (Table 4) and 0 otherwise. "Mispricing" and "Post-2004" are defined similarly. Parentheses show standard errors clustered by month. "Null: Risk No Decay" shows the p-value that tests whether risk-based returns do not decrease post-sample ((1) and (3)) or post-publication ((2) and (4)). Risk-based predictors decay more than other predictors, but the difference is only marginally significant. The decay in risk-based predictors overall is highly significant.

| | LHS: Long-Short Strategy Return (bps pm, scaled) | | | | |
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100.0	100.0	100.0	100.0	102.4
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	-42.3	-25.3	-36.5	-24.4	0.7
	(8.6)	(11.7)	(10.3)	(15.3)	(14.5)
Post-Pub		-21.0		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	-28.7	-18.5	-34.4	-19.5	-23.4
	(15.4)	(20.2)	(17.1)	(22.8)	(15.2)
Post-Pub x Risk		-14.2		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8.1	-1.3	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-8.7	
				(17.5)	
Post-2004					-59.6
					(16.6)
Null: Risk No Decay	< 0.1%	< 0.1%	< 0.1%	< 0.1%	< 0.1%

Additional robustness is shown in specification (5), which controls for the idea that information technology has led to weaker predictability post-2004 (Chordia, Subrahmanyam, and Tong (2014); Chen and Velikov (2022)). In this specification, risk-based predictors still decay by an additional 23 percentage points. We also find similar results using regressions without normalizing returns (Appendix Table A.6).

A more refined control for time effects is found in Figure 4. As in Figure 1, we construct data-mined benchmarks by searching 29,000 accounting signals for t-stats > 2.0 in the risk-based predictors' original sample periods. These benchmark returns are thus

exposed to the same time effects as the risk-based predictors, such as business cycles and interest rate regimes.

Figure 4: Risk-Based Predictors vs Data-Mined Benchmarks

'Published' includes only predictors that are based on risk according to peer review (Table 4). '$|t| > 2.0$ Mining Accounting' is a data-mined benchmark formed by filtering 29,000 strategies for $|t| > 2.0$ in published predictors' original sample periods. The plot shows long-short returns in event time, where the event is the end of the original sample periods. All predictor returns are normalized to average 100 bps in the original samples. Risk-based predictors underperform data mined predictors that are exposed to the same time effects.
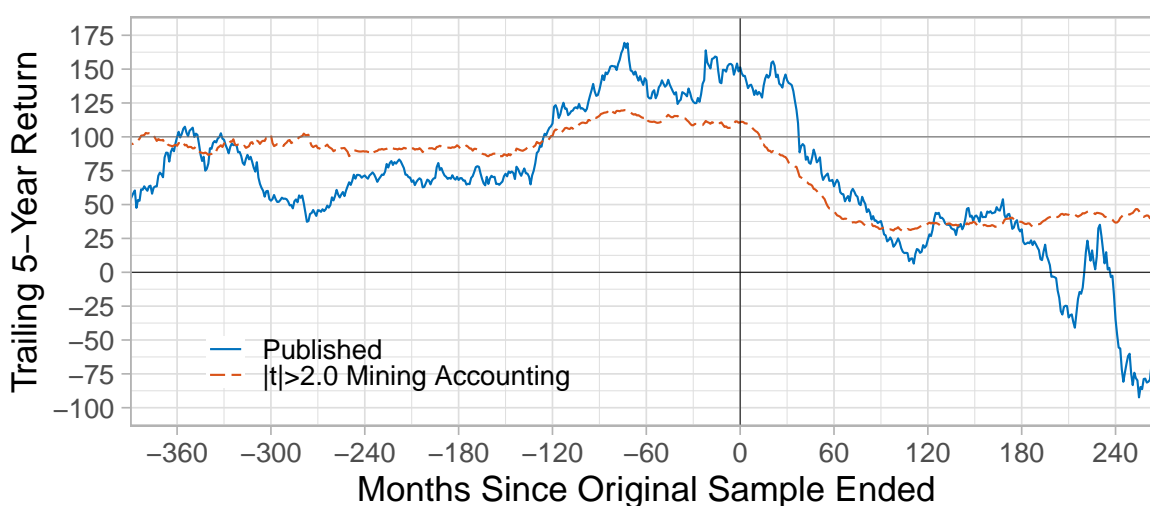
Controlling for time effects in this way, risk-based predictors still underperform. The trailing 5-year returns (solid) are not fully post-sample until 60 months after the original samples. Trailing returns in this region are at or below the data-mined benchmarks (dashed line). The analogous plots for mispricing-based and agnostic predictors are in Appendix Figure A.1, and show that agnostic predictors slightly outperform, an issue we return to in Section 5.1.

Taken together, these results demonstrate an important asymmetry in how investors learn from academic publications. While it seems that investors learn about mispricing (McLean and Pontiff (2016)), they do not seem to learn about risk. If they did learn about risk, they would buy the safe stocks and sell the risky ones, increasing predictability post-publication. Such an increase is strongly rejected by Table 5. This broken symmetry implies that either peer review systematically mislabels mispricing as risk or that peer review on average finds unstable risk that decays over time. Neither implication is positive for the peer review process.

4.3   Do Mathematical Models Help?

A common belief is that theory protects against post-sample decay by restricting the number of possible signals (e.g. Harvey, Liu, and Zhu (2016)).[5] Perhaps the risk-based predictors in Section 4.2 are not restricted enough. In fact, many of the risk-based predictors are supported by informal arguments rather than rigorous equilibrium theory. Does focusing on predictors supported by mathematical models help?

To address this question, we categorize predictors by the rigor of the mathematical model (if any) that is used as supportive evidence. The categories we consider are stylized model (e.g. a two-period model), dynamic equilibrium (many periods), or quantitative equilibrium (calibrated to match key moments in the data). We then examine the post-sample performance by model type. As with our risk-vs-mispricing categorizations, our rigor categories are public at `https://github.com/chenandrewy/flex-mining/blob/main/DataInput/SignalsTheoryChecked.csv`.

Figure 5 shows the result. 25 of the 199 predictors in our sample are supported by a mathematical model, and only 6 of these are quantitative equilibrium. While this sample is small, point estimates imply the opposite of what is commonly believed. The mean normalized post-sample return is monotonically *decreasing* in modeling rigor.

This result is notable given the impressive work behind quantitative equilibrium models. This work is not just theoretical: it is also computational and empirical. 'realestate' is based on Tuzel's (2010) general equilibrium production economy with heterogeneous firms. Though this class of models is difficult to solve, Tuzel does not simplify for tractability, and instead computes the equilibrium using Krusell and Smith (1998) approximate aggregation and the parameterized expectations algorithm of Marcet (1991). She then calibrates the equilibrium to match many moments in the data. The calibration shows that her model is not just a qualitative description, but that it is a quantitative match for the U.S. economy. Despite all of this rigor, 'realestate' returned only 1 bps per month between the end of Tuzel's sample period (2005) and the end of our dataset (2022).
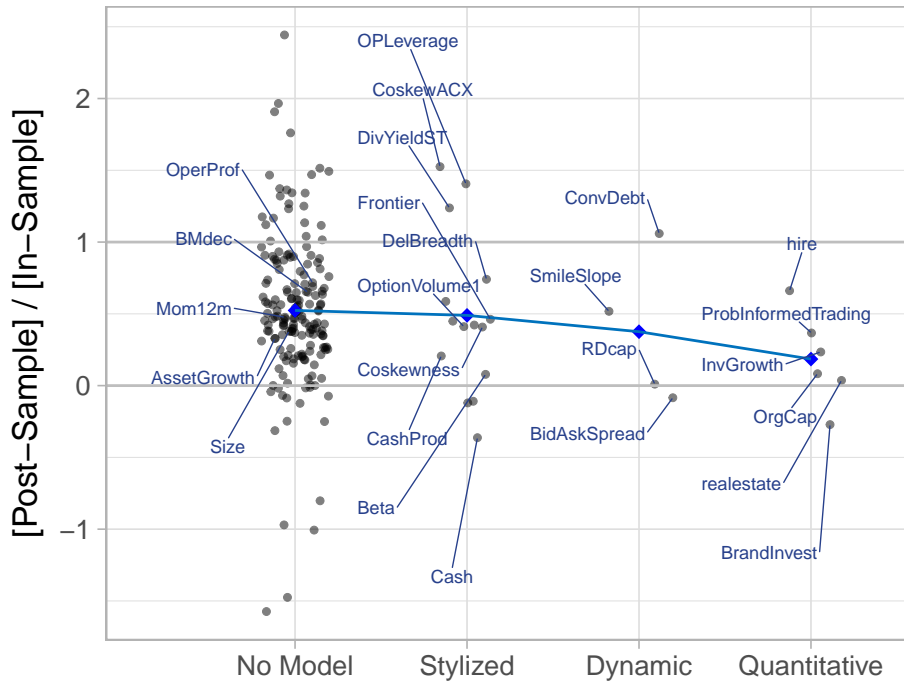
The other quantitative equilibrium models in Figure 5 are not general equilibrium, but they are still impressive. Each one solves a dynamic or asymmetric information model numerous times to ensure that the selected model matches many features of real world data. This work leads to predictors that under perform papers that use stylized models or informal arguments.

One interpretation is that a talented theorist can justify nearly any empirical pattern (the Sonnenschein-Mantel-Debreu "Anything Goes Theorem," Mas-Colell, Whinston,

---

[5]The meta-theory in Section 4.4 implies this belief is misguided.

Figure 5: Post-Sample Returns by Model Rigor

Each circle represents one published predictor's post-sample mean return normalized by its original-sample return. 'Stylized,' 'dynamic,' and 'quantitative' are the type of models used as supporting evidence for the predictor, with 'quantitative' models being dynamic or asymmetric information models calibrated to match important moments in the data. Diamonds show means across predictor by category. The reference for each acronym can be found at `https://github.com/chenandrewy/flex-mining/blob/main/DataInput/SignalsTheoryChecked.csv`. The estimated relationship between modeling rigor and post-sample performance is monotonic and negative.

and Green (1995)), especially with the aid of modern computing power. If this is the case, then the support of a quantitative equilibrium model may not provide any additional information about the predictor. Another interpretation is that the set of theories is restricted (perhaps by the judgement of peer review) but that these restrictions are unrelated to the risks of concern to real-world investors. We formalize the latter interpretation in Section 4.4.

Regardless of the interpretation, these results raise questions about the idea that economic priors should be imposed in the peer review process (Harvey, Liu, and Zhu (2016); Harvey (2017)). At least some peer reviewers would have the prior that quantitative equilibrium models are a sign of out-of-sample robustness. Figure 5 suggests that such priors need updating.

4.4 A Model of the "Cochrane Conjecture"

Our results conflict with the common belief that theory protects against data mining bias. We call this belief the "Cochrane Conjecture," after the discussion of this issue in Cochrane's (2009) textbook. This section provides a formal analysis of this conjecture and provides sufficient conditions for its validity. The model is a slight generalization of models in Chen and Zimmermann (2020) and Chen and Velikov (2022).

The long-short return of predictor i in month $t = 1, 2, \ldots$ depends on whether t is in the original sample ($t \leq T_i$) or post-sample ($T_i < t \leq T_i + T_i^{OOS}$):

$$
r_{i,t} = \begin{cases} \mu_i + \varepsilon_{i,t} & t \leq T_i \\ \mu_i + \Delta\mu_i + \varepsilon_{i,t} & T_i < t \leq T_i + T_i^{OOS} \end{cases}.
\tag{1}
$$

where $\mu_i \equiv E(r_{i,t} | t \leq T_i)$ is the expected return in the original sample, $\Delta\mu_i \equiv E(r_{i,t} | T_i < t \leq T_i + T_i^{OOS}) - \mu_i$ is the change in expected returns change post-sample, and $\varepsilon_{i,t}$ is a zero mean residual. Sample mean returns during the original and post-sample periods are denoted by $\bar{r}_i \equiv T_i^{-1} \sum_{t=1}^{T} r_{i,t}$ and $\bar{r}_i^{OOS} \equiv (T_i^{OOS})^{-1} \sum_{t=T_i+1}^{T_i+T_i^{OOS}} r_{i,t}$, respectively. The post-sample decay is defined as

$$
\text{Decay}_i \equiv \text{Sign}(\bar{r}_i) \left(\bar{r}_i - \bar{r}_i^{OOS} \right),
\tag{2}
$$

where $\text{Sign}(\bar{r}_i)$ flips the long and short legs if the original sample mean return is negative. Here, a positive Decay_i means that predictability weakens post-sample. These expressions are definitions rather than economic restrictions.

The only economic restriction is that post-sample residuals are unpredictable using original sample information:

$$
E \left(\varepsilon_{i,t} \,\middle|\, \mathcal{I}_i^{IS}, T_i < t \leq T_i + T_i^{OOS} \right) = 0
\tag{3}
$$

where $\mathcal{I}_i^{IS}$ represents any information available at month T_i for predictor i (e.g. the original sample mean return $\bar{r}_i$).

Data mining involves searching through a set $\mathcal{D}$ of predictors (e.g. the accounting-based predictors in Section 2.1) and selecting those with $|\bar{r}_i|$ above some threshold h. One can show the expected post-sample decay from this process is

$$
E \left(\text{Decay}_i \mid |\bar{r}_i| > h, i \in \mathcal{D} \right) = E \left(\text{Sign}(\bar{r}_i)(-\Delta\mu_i + \bar{\varepsilon}_i) \mid |\bar{r}_i| > h, i \in \mathcal{D} \right).
\tag{4}
$$

The post-sample residuals vanish due to Equation (3), but the original sample's mean residual $\bar{\varepsilon}_i$ remains, due to the selection on $\bar{r}_i > h$. Thus, there are two reasons for post-sample decay: either the fundamental expected return shifts toward zero ($\text{Sign}(\bar{r}_i)\Delta\mu_i < 0$) or the selection process picks up lucky original sample reutrns ($\text{Sign}(\bar{r}_i)\bar{\varepsilon}_i > 0$).

Using economic theory amounts to searching through a different set $\mathcal{T}$ (e.g. risk-based predictors). The question is: under what conditions does searching through $\mathcal{T}$ lead to less decay? The following proposition provides sufficient conditions:

Proposition 1. *(Cochrane Conjecture) If the following inequalities hold:*

$$E\left(Sign(\bar{r}_i)\mu_i \mid |\bar{r}_i|, i \in \mathcal{T}\right) > E\left(Sign(\bar{r}_i)\mu_i \mid |\bar{r}_i|, i \in \mathcal{D}\right) \tag{5}$$

$$E\left(Sign(\bar{r}_i)\Delta\mu_i \mid |\bar{r}_i|, i \in \mathcal{T}\right) > E\left(Sign(\bar{r}_i)\Delta\mu_i \mid |\bar{r}_i|, i \in \mathcal{D}\right), \tag{6}$$

then theory leads to less decay than data mining:

$$E\left[Decay_i \mid |\bar{r}_i| > h, i \in \mathcal{T}\right] < E\left[Decay_i \mid |\bar{r}_i| > h, i \in \mathcal{D}\right].$$

The proof is in Appendix A.4.

Proposition 1 provides a formal justification for the argument in Chapter 7 of Cochrane (2009). There, Cochrane describes the problem of "dredging up spuriously good in-sample pricing," and argues that "the best hope for finding pricing factors that are robust out of sample... ...is to try to understand the fundamental macroeconomic sources of risk."

As implied by the quote, improved robustness is not guaranteed, but it obtains under some conditions. According to Proposition 1, one condition is that understanding risk leads to higher expected returns than data mining—holding fixed summary statistics about past performance (Equation (5)). Another condition is that understanding risk leads to more stable expected returns—once again holding fixed summary statistics (Equation (6)). The importance of risk and equilibrium is implicit in these expressions. Unlike predictability that is based on mispricing, risk-based predictability should be stable both within the original sample ($\bar{r}_i$ is due to μ_i and not $\bar{\varepsilon}_i$) and post-sample ($\Delta\mu_i$ does not shift expected returns toward zero). If both of these conditions hold, then Proposition 1 says improved out-of-sample robustness *must* obtain.

Our empirical results imply that at least one of these conditions fails to hold for the understanding of risk embodied by the papers in the Chen and Zimmermann (2022) dataset. Either cause is consistent with a disconnect between the risks identified in these papers and the risks that are of concern to real-world investors. This disconnect is found in many surveys of real-world investors, from finance professionals (Mukhlynina

and Nyborg (2020); Chinco, Hartzmark, and Sussman (2022)), to millionaires (Bender et al. (2022)), to tenured finance professors (Doran and Wright (2007)). In all of these surveys, risks that are important according to peer review are unimportant for real world decisions.

Notably, the size of the sets $\mathcal{T}$ and $\mathcal{D}$ does not appear in Proposition 1. In other words, the *amount* of data mining is irrelevant, in contrast the common intuition that more tests imply more false discoveries (e.g. Harvey, Liu, and Zhu (2016)). Instead, what matters is whether theory or data mining provides a better signal of the fundamental expected returns (μ_i and $\Delta\mu_i$). This theory of data mining is consistent with our finding that data-mined ticker predictors decay more than accounting predictors (Figure 2)—despite the fact that the ticker-based predictors are mined from a much smaller dataset.

5 Robustness

This section demonstrates robustness. We control for original-sample mean returns, t-stats, and correlations (Section 5.1), exclude predictors that are correlated with all existing research (Section 5.2), focus on the most well-studied predictors (Section 5.3), and try alternative measures of risk (Section 5.4).

5.1 Controlling for Sample Mean Returns, t-stats, Correlations

The previous results may be unfair to peer-reviewed predictors, as the data-mined predictors may have stronger statistical support. It would not be fair, for example, to compare data-mined predictors with t-statistics of 6.0 a peer-reviewed predictor with a t-statistic of 2.0. One may also be concerned about that normalizing mean returns to be 100 bps per month somehow biases the results.

To account for these issues, we use a more restrictive matching procedure. For each published predictor, we match with data-mined predictors that have t-statistics within 10% and mean returns within 30% of the published predictors, using the original sample periods. Appendix Figure A.4 shows similar results using the 10% threshold for mean returns. As in Section 3, we also restrict the data-mined predictors to match the published ones in terms of equal- or value-weighting. We then repeat our primary post-sample tests (Figures 1 and 4).

Figure 6 shows the result. Data mining continues to perform similarly to peer review overall (Panel (a)) and outperforms risk-based predictors out-of-sample (Panel (b)). This results is perhaps natural, as it is unlikely that data mining tends to uncover larger

t-statistics than peer review, since researchers themselves can use data mining.

Panels (c) and (d) look closer at mispricing and agnostic predictors. Data mining closely mimics the returns of data-mined predictors but it somewhat underperforms agnostic predictors. These results are consistent with data-mining benchmarks that simply screen on t-stats > 2.0 (Appendix Figure A.1), so it is not the more restricted screen that generates the outperformance of agnostic predictors. A potential explanation for this outperformance is that many of the agnostic predictors are based on past returns. Past-return predictors are missing from the accounting ratios we use for the data mining benchmarks (Section 2.1). Several agnostic past-return predictors have performed quite well post-sample (e.g. Moskowitz and Grinblatt's industry momentum (1999) and several versions of Heston and Sadka's seasonal momentum (2008)).

Another potential concern is that correlations may be driving our results. This idea can be motivated by the idea that expected returns are driven by correlations with risk factors.

To address this concern, we additionally exclude data-mined strategies that have correlations of more than 0.10 with the target published strategy in the original samples. This additional filter (red long-dashed line) has little effect on the path of 5-year returns in any of the four panels of Figure 6.

This robustness is natural given the diversity of data-mined predictors from Table 2. The majority of data-mined predictors with t-stats > 2.0 have correlations less than 0.25 in absolute value.

5.2   Controlling for Correlation with All Existing Research

The previous analysis excludes data-mined benchmarks that are correlated with the published predictor in question. However, it may still be the case that the included benchmarks are spanned by other previously-published predictors or linear combinations of them. This subsection examines these possibilities.

Before we begin, we point out that data mining uncovers academic themes long before they are published (Table 3). So, if the question were reversed, and one were to ask whether published predictors are spanned by data-mined ones, the answer would be a solid yes. Moreover, the peer review process seems to uncover data-mined predictors with the very largest t-stats (Figure 2, Panel (b)). Empirical Bayes logic, then, implies that excluding these data-mined predictors will lead to benchmarks with lower post-sample returns (Chen and Zimmermann (2020)).

Panel (a) of Figure 7 re-examines Figure 1, but separates the data-mined benchmarks

Figure 6: Controlling for Sample Mean Returns, t-stats, Correlations

We compare published strategies to data mined benchmarks based on original-sample t-stats (as in Figures 1 and 4) but now we drop data-mined strategies if they have t-stats that differ by more than 10% or mean returns that differ by more than 30% (short-dash). We additionally drop data-mined strategies that are more than 10% correlated with published strategies in the original sample (long-dash). The similarity in post-sample returns is not driven by very high data-mined t-stats, the normalization of returns to be 100 bps in sample, nor by correlation with the published strategy.


(a) All

(b) Risk-based

(c) Mispricing-based

(d) Agnostic

into those that have a correlation > 0.50 with any existing published predictor and those that do not. Here we define existing published predictors as those that use samples ending at the same time or earlier than the published predictor in question. We further separate the low correlation benchmarks into those with t-stats higher than the published predictor and those with lower t-stats.

Generally speaking, all three types of data-mined benchmarks perform similarly to the

Figure 7: Excluding Data-Mined Benchmarks Correlated with Any Existing Research

We compare published predictors (solid) to data-mined benchmarks made with t-stats >
2.0, but separate benchmarks into high correlation (long-dash), low correlation with high
in-sample t-stats (short-dash), and low correlation with low in-sample t-stats (dot-dash).
Panel (a) uses the maximum pairwise correlation with any existing published predictor,
where existing publications are those with samples ending at the same time or earlier than
the published predictor in question. Panel (b) uses the R^2 from regressing the data-mined
return on 5 principal components of existing predictors computed using probabilistic
PCA. Data-mined benchmarks are generally similar to published predictors, regardless of
the correlation. High correlation benchmarks outperform, consistent with the fact that
publication selects for the data-mined predictors with the very highest t-stats (Figure
2). 5-year returns co-move across all groups, consistent with aggregate movements in
mispricing (Table 3; Stambaugh, Yu, and Yuan (2012); Chen and Velikov (2022)).

(a) Any Correlation   (b) PPCA
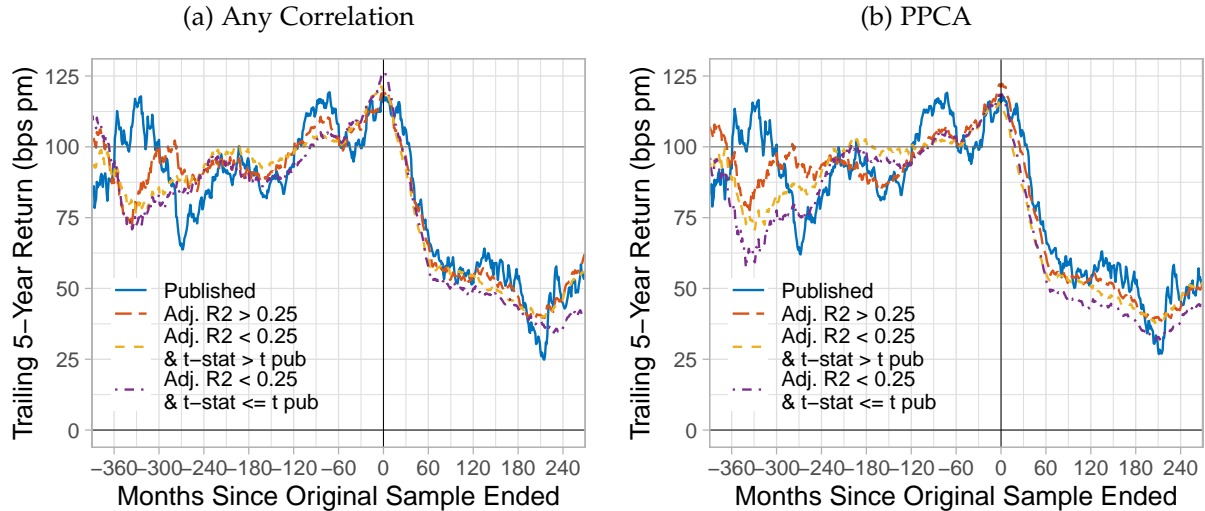
published predictors. As implied by empirical Bayes and Figure 2, excluding correlated predictors leads to lower post-sample returns (dash-dot line). Nevertheless, the low correlation benchmarks that have higher t-stats perform quite similarly to published predictors throughout the post-sample period (short-dash line).

Panel (b) of Figure 7 measures correlations with five factors extracted from existing publications via probabilistic principal component analysis (PPCA, Roweis (1997)). This estimate accounts for spanning not only with individual existing publications but linear combinations of them. PPCA is a natural way to both handle missing data and regularize these estimates (Chen and McCoy (2024)).

We find PPCA leads to similar results. All benchmarks perform similarly to published strategies. The low correlation benchmarks generally lead to somewhat lower post-sample returns, but the low correlation benchmarks with higher t-stats perform quite similarly to published predictors.

An interesting feature throughout Figure 7 is that the trailing 5-year returns are highly correlated, even across groups of returns that have low monthly return correlations. This result is unlikely to be due to autocorrelation in the transient component of monthly returns: the median autocorrelation in monthly returns across data-mined predictors is 0.07 and the 95th percentile is 0.18.

Instead, the co-movement in 5-year returns is likely due to correlation in the long-run expected return component. As seen in Table 3, the expected returns of many data-mined predictors decays post-2004 (see also Appendix Table A.1). Moreover, Yan and Zheng (2017) find that the expected returns of data-mined predictors is positively correlated with aggregate investor sentiment. This co-movement in expected returns is also found among published predictors (Stambaugh, Yu, and Yuan (2012); Chen and Velikov (2022)), and is consistent with mispricing being a key driver of cross-sectional predictability.

5.3   B/M and Momentum vs Data Mining

Perhaps the very best research produces predictors that out-perform data mining. To examine this possibility, we take a closer look at the two most renown predictors in the literature: B/M and momentum. These predictors are not only the most famous, but arguably the ones with the most well-documented supporting evidence, both theoretical (e.g. Gomes, Kogan, and Zhang (2003); Campbell and Vuolteenaho (2004); Hong and Stein (1999); Daniel, Hirshleifer, and Subrahmanyam (1998)), and empirical (e.g. Fama

and French (1993); Asness, Moskowitz, and Pedersen (2013)).[6]

Table 6 compares Fama and French's (1992) version of B/M with 171 data-mined benchmarks that have mean returns within 30% and t-stats within 10% of BM's statistics, using Fama and French's 1963-1990 sample period. It lists 20 of the 171 predictors. The benchmarks are ranked by their similarity with B/M in terms of mean returns.

The data-mined benchmarks include themes that have been found in the cross-sectional literature, such as asset growth, issuance, and accruals—themes that were documented *after* Fama and French (1992). On average, the 171 benchmarks earned 83 bps per month in the 1963-1990 sample, a touch below the 96 bps per month earned by B/M. Post 1991, the 171 benchmarks earned on average 69 bps per month, outperforming B/M by 7 bps.

Table 7 applies the same exercise to Jegadeesh and Titman's (1993) 12-month momentum. Since momentum has a much higher mean return, only 44 data-mined benchmarks are found. Some of themes from Table 6 show up again in Table 7, though we see some unusual variables like rental expense, depreciation, and income taxes. Unlike with B/M, here peer review outperforms somewhat, with momentum earning 66 bps per month post-sample compared to 48 bps for data mining.

Though the samples are small, Tables 6 and 7 suggest that focusing on the most renown predictors does not significantly affect our results. These two predictors that academics deem to be most worthy of attention perform similarly to data mining post-sample.

5.4 Alternative Measures of Risk

The previous measures of risk are binary: a predictor is either risk-based or not. But predictors may be due to a mixture of risk and mispricing. Perhaps a more continuous measure of risk could help predict post-sample returns.

To examine this possibility, Figure 8 plots post-sample returns against the ratio of risk words (e.g. "utility," "equilibrium") to mispricing words (e.g. "sentiment," "underreact") in the published papers. The risk and mispricing words are counted by software and defined in Appendix A.1.

The figure shows a negative relationship between post-sample returns and the ratio of risk to mispricing words. Thus, the underperformance of risk predictors is not due to an artifical binary classification. This result is also consistent with the monotonically

---

[6]Other papers that provide theoretical support for B/M include Zhang (2005); Lettau and Wachter (2007); Gabaix (2008); Papanikolaou (2011); and Chen (2018). Other papers that provide theoretical support for momentum include Brav and Heaton (2002); Holden and Subrahmanyam (2002); and Da, Gurun, and Warachka (2014). For a recent review of momentum theories see Subrahmanyam (2018).

Table 6: 20 Data-Mined Predictors With Returns Similar to Fama-French's B/M (1992)

Table lists 20 of the 171 data-mined signals that performed similarly to Fama and French's (1992) B/M in the original 1963-1990 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. investment, equity issuance, accruals) and leads to similar out-of-sample performance as Fama and French's B/M.

Similarity Rank	Signal	Sign	Mean Return (% p.m.) 1963-1990	1991-2021
Peer-Reviewed				
	Book / Market (Fama-French 1992)	1	0.96	0.62
Data-Mined				
1	Δ[Assets]/lag[Operating expenses]	-1	0.96	0.90
2	Δ[PPE net]/lag[Sales]	-1	0.96	0.80
3	[Market equity FYE]/[Depreciation & amort]	-1	0.95	0.66
4	Δ[Assets]/lag[Cost of goods sold]	-1	0.95	0.86
5	Δ[Assets]/lag[SG&A]	-1	0.95	0.82
6	Δ[PPE net]/lag[Gross profit]	-1	0.98	0.51
7	Δ[PPE net]/lag[Current liabilities]	-1	0.94	0.90
8	[Depreciation (CF acct)]/[Capex PPE sch V]	1	0.97	0.78
9	[Market equity FYE]/[Depreciation depl amort]	-1	0.94	1.03
10	[Stock issuance]/[Capex PPE sch V]	-1	0.94	0.93
	. . .			
101	[Market equity FYE]/[Current assets]	-1	1.15	0.89
102	[Market equity FYE]/[Common equity]	-1	1.14	0.51
103	Δ[Cost of goods sold]/lag[Cost of goods sold]	-1	0.75	0.94
104	Δ[Receivables]/lag[Invested capital]	-1	0.76	0.46
105	Δ[Receivables]/lag[PPE net]	-1	0.76	0.46
	. . .			
167	Δ[Assets]/lag[IB adjusted for common s]	-1	0.67	-0.02
168	Δ[Cost of goods sold]/lag[Current liabilities]	-1	0.67	0.68
169	Δ[Long-term debt]/lag[Operating expenses]	-1	0.67	0.50
170	Δ[Depreciation & amort]/lag[Invested capital]	-1	0.67	0.68
171	Δ[Current liabilities]/lag[Inventories]	-1	0.67	0.49
	Mean Data-Mined		0.83	0.69

negative relationship between model rigor and post-sample performance (Figure 5).

One can alternatively measure risk using factor models, as follows. For each published long-short portfolio i, we estimate exposure to factor k using time-series regressions on the original papers' sample periods. According to the factor models, the estimated

Table 7: 20 Data-Mined Predictors That Perform Similarly to Jegadeesh and Titman's Momentum (1993)

Table lists 20 of the 44 data-mined signals that performed similarly to Jegadeesh and Titman's (1993) 12-month momentum in the original 1964-1989 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. profitabitability, investment) and leads to similar out-of-sample performance as Jegedeesh and Titman's momentum.

Similarity Rank	Signal	Sign	Mean Return (% p.m.)	
			1964-1989	1990-2021
Peer-Reviewed				
	12-Month Momentum (Jegadeesh-Titman 1993)	1	1.38	0.66
Data-Mined				
1	[Retained earnings unadj]/[Market equity FYE]	1	1.38	-0.19
2	[Retained earnings unadj]/[Assets other sundry]	1	1.40	0.15
3	[PPE and machinery]/[Current liabilities]	1	1.42	0.38
4	[Retained earnings unadj]/[Cash & ST investments]	1	1.42	0.17
5	[PPE and machinery]/[Capital expenditure]	1	1.50	0.79
6	[Retained earnings unadj]/[Invest & advances other]	1	1.51	0.03
7	[Investing activities oth]/[Nonoperating income]	1	1.52	0.10
8	[Income taxes paid]/[PPE net]	1	1.22	0.14
9	[PPE and machinery]/[Capex PPE sch V]	1	1.56	0.80
10	[Market equity FYE]/[Current assets]	-1	1.20	0.88
	. . .			
21	[Depreciation (CF acct)]/[Market equity FYE]	1	1.10	0.72
22	[Retained earnings unadj]/[Common equity]	1	1.66	-0.08
23	Δ[Assets]/lag[Current assets]	-1	1.09	1.26
24	Δ[PPE (gross)]/lag[Operating expenses]	-1	1.09	0.70
25	[Funds from operations]/[Market equity FYE]	1	1.07	-3.51
	. . .			
40	[Deprec end bal (Sch VI)]/[Market equity FYE]	1	1.02	1.03
41	[Market equity FYE]/[Cost of goods sold]	-1	1.00	0.70
42	[Rental expense]/[Market equity FYE]	1	1.00	0.84
43	Δ[Invested capital]/lag[Current assets]	-1	0.97	1.32
44	[Retained earnings unadj]/[Receiv current other]	1	1.79	0.19
	Mean Data-Mined		1.29	0.48

expected return is $\sum_k \hat{\beta}_{k,i} \bar{f}_k$, where $\bar{f}_k$ is the original-sample mean return of factor k. Fama and French (1993) state that $\hat{\beta}_{i,k}$ with respect to their SMB and HML factors have "a clear interpretation as risk-factor sensitivities." If this interpretation is both correct and stable,

Figure 8: Post-Sample Returns vs Risk to Mispricing Words

Each marker represents one published predictor's mean return. The regression line is fitted with OLS. The full reference for each acronym can be found at `https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv`. The relationship between risk words and post-sample returns is negative.
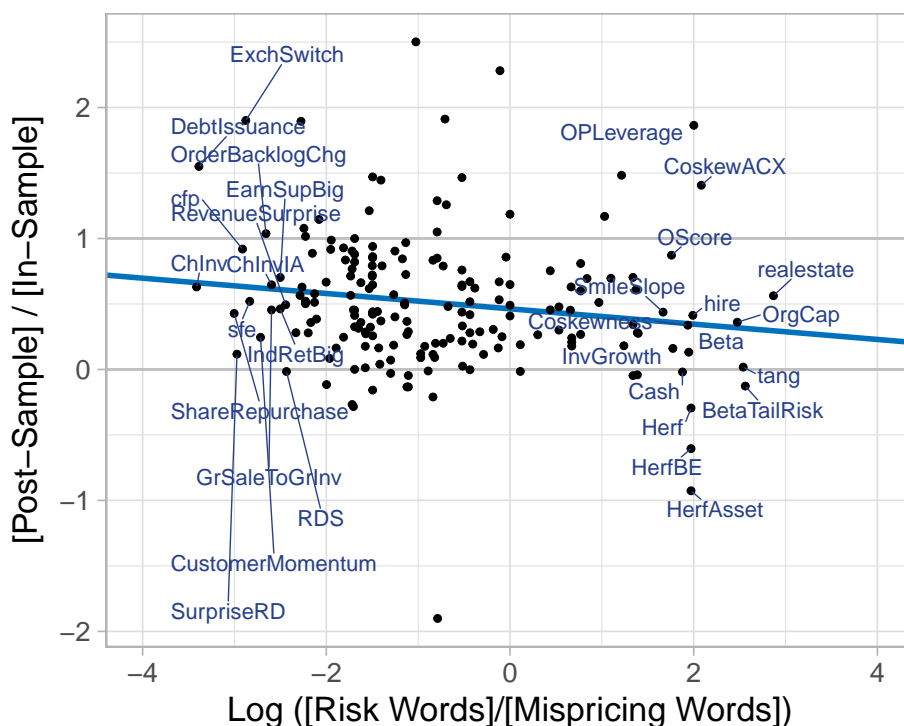
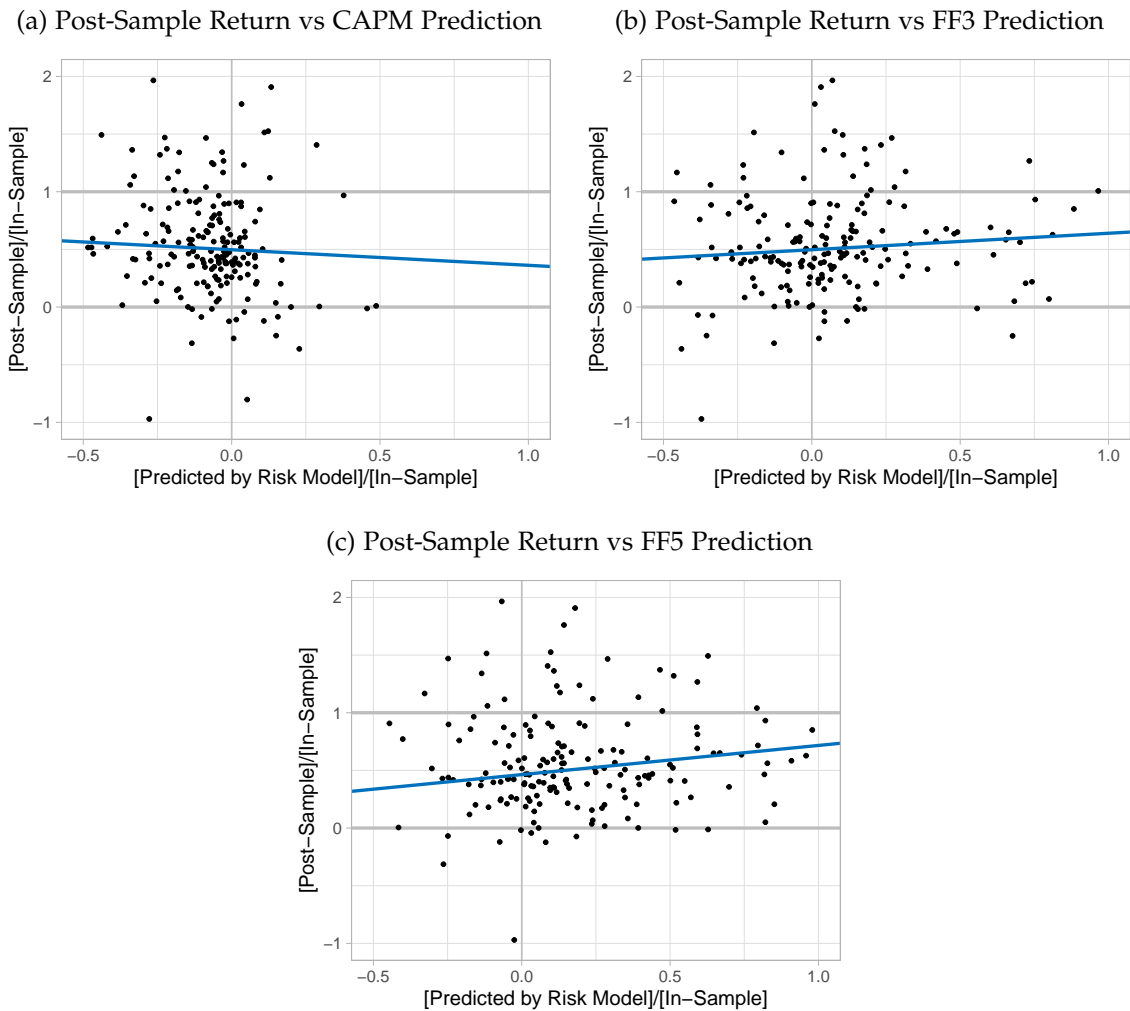then the estimated expected return should remain post-sample.

Figure 9 plots the post-sample mean return against the factor model expected returns, using the CAPM, Fama-French 3 (FF3), or Fama-French 5 (FF5) models. We normalize by the original-sample mean return for ease of interpretation. With this normalization, the position on the x-axis ([Predicted by Risk Model]/[In-Sample]) represents the share of predictability due to risk.

The figure shows that a minority of in-sample predictability is attributed to risk, at best. Using the CAPM (Panel (a)), nearly all predictability is less than 25% due to risk (to the left of the vertical line at 0.25), and many predictors have a *negative* risk share. FF3 (Panel (b)) implies more predictability is due to risk, but still the vast majority of predictors lie to the left of 0.50.

Fama and French (2015) are more cautious than Fama and French (1993), and describe the risk-based ICAPM as "the more ambitious interpretation" of the five factor model. Under the more ambitious interpretation, FF5 implies that most predictors are less than

Figure 9: Mean Returns Post-Sample vs Factor Model Predictions

Each marker is one published long-short strategy. [Post-Sample]/[In-Sample] is the mean return post-sample divided by the mean return in-sample. [Predicted by Risk Model] is $\sum_k \hat{\beta}_{k,i} \bar{f}_k$, where $\bar{f}_k$ is the in-sample mean return of factor k and $\hat{\beta}_{k,i}$ comes from an in-sample time series regression of long-short returns on factor realizations. FF3 and FF5 are the Fama-French 3- and 5-factor models. The blue line is the OLS fit. The axes zoom in on the interpretable region of the chart and omits outliers. Factor models attribute a minority of in-sample predictability to risk, at best. Post-sample decay is the distance between the horizontal line at 1.0 and the regression line, and this decay is near 50% even for predictors that are entirely due to risk according to the CAPM and FF3. For FF5, decay is smaller for predictors that are more than 75% due to risk, but these predictors are rare.

(a) Post-Sample Return vs CAPM Prediction

(b) Post-Sample Return vs FF3 Prediction

(c) Post-Sample Return vs FF5 Prediction

50% due to risk. These results are consistent with our manual reading of the papers, which typically attribute predictability to mispricing (Table 4).

The regression lines in Figure 9 show negative or mildly positive relationships between

factor model risk and post-sample returns. The regression fits for the CAPM and FF3 models never stray far from 50%, implying that even predictors that are entirely due to risk are little different than the typical predictor in terms of post-sample robustness. FF5 risk shows a stronger relationship with post-sample returns, but even the rare predictors that are 75% due to risk decay by roughly 40% post-sample. Moreover, the Fama and French (2015) model may have the benefit of hindsight, as the median publication year for the Chen and Zimmermann (2022) predictors is 2006.

6 Conclusion: a "bitter lesson" for asset pricing?

Sutton (2019) reflects on 70 years of artificial intelligence research, in areas ranging from chess to computer vision to natural language processing. He arrives at "the Bitter Lesson": Time after time, hand-crafted solutions end up "irrelevant, or worse," while vast searches through big datasets outperform. His broader takeaway is that actual minds are "tremendously and irredeemably complex," as is the real world they inhabit.

We document a kind of "bitter lesson" for asset pricing. We show that mining tens of thousands of accounting ratios for statistical significance leads to post-sample returns comparable to the peer review process. Peer-reviewed risk theories do not help, and in fact predictors supported by such theories *under*perform. More rigorous theory does not help either. If anything, more rigor leads to worse post-sample performance.

Asset pricing is not chess nor computer vision. The data are smaller, noisier, and subject to arbitrage forces (Kelly, Israel, and Moskowitz (2020)). In our setting, it is not at all clear that vast searches through big datasets will always outperform.

But Sutton's broader takeaway may be the underlying driver of our findings. Could it be that the minds of investors, and the firms they aim to price, are also "tremendously and irredeemably complex?" Complexity is evident in the thousands of predictors that we document, as well as the dozens of principal components required to span them. Complexity is also consistent with the fact that the elegant and parsimonious theories sought after by finance scholars (Cochrane (2017)) have failed to out-predict sheer data mining. This failure comes despite the decades of efforts embodied by the Chen and Zimmermann (2022) meta-study, and the high-powered incentives of finance academia (Celerier, Vallee, and Vasilenko (2022)).

Regardless of the underlying driver, a clear takeaway is that data mining is undervalued in asset pricing. Data mining uncovers out-of-sample predictability, as strong as is uncovered by the best minds in finance. And while data mining introduces bias, multiple testing methods can remove this bias, if done correctly (Chen and Zimmermann (2023);

Chen and Dim (2023)).

We do not argue economists should become engineers and abandon elegant and parsimonious theories. Such theories are our strength—they are the very meaning of the expression "the economics." Instead, we argue that data mining could be the key to ensuring our theories stay relevant. Put another way, completely exploring the data should lead to theories that are closer to the fundamental sources of returns of the real world.

Appendix A   Appendix

A.1   Risk words and mispricing words

We remove stopwords, lowercase and lemmatize all words using standard methods. Then, we count separately the words corresponding to risk and mispricing.

We consider as risk words the following terms and their grammatical variations: "utility," "maximize," "minimize," "optimize," "premium," "premia," "premiums," "consume," "marginal," "equilibrium," "sdf," "investment-based," and "theoretical." We also count as risk words appearances of "risk" that are not preceded by "lower," and appearances of "aversion," "rational," and "risky" that are not preceded by "not."

The mispricing words consist of " "anomaly," "behavioral," "optimistic," "pessimistic," "sentiment," "underreact," "overreact," "failure," "bias," "overvalue," "misvalue," "undervalue," "attention," "underperformance," "extrapolate," "underestimate," "misreaction," "inefficiency," "delay," "suboptimal," "mislead," "overoptimism," "arbitrage," "factor unlikely," and their grammatical variations. We further count as mispricing the terms "not rewarded," "little risk," "risk cannot [explain]," " low [type of] risk," "unrelated [to the type of] risk," "fail [to] reflect," and "market failure," where the terms in brackets are captured using regular expressions or correspond to stopwords.

A.2 Robustness: Data-Mined Predictability

Table A.1: out-of-sample Returns from Mining Accounting Data: 2004-2020

We sorts 29,000 data-mined strategies each June into 5 bins based on past 30-year mean returns (in-sample) and computes the mean return over the next year within each bin (out-of-sample). Statistics are calculated by strategy, then averaged within bins, then averaged across sorting years. Decay is the percentage decrease in mean return out-of-sample relative to in-sample. We omit decay for bin 4 because the mean return in-sample is negligible. Post-2004, out-of-sample returns are much weaker, though they still exist.

In-	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
Sample	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
Bin	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.9	-4.04	-27.1	54.8	-37.2	-1.88	-4.6	87.6
2	-28.4	-2.32	-10.6	62.7	-14.4	-0.90	-0.7	95.1
3	-11.6	-1.00	-0.1	99.1	-3.8	-0.25	-1.7	55.5
4	2.3	0.19	7.8		6.1	0.40	-3.3	
5	25.9	1.53	17.4	32.8	28.4	1.38	2.3	91.9

Table A.2: Themes from Mining Accounting Ratios in 1990

Table reports the 20 accounting ratio numerator and stock weight (equal- or value-) combinations with the largest mean t-stats using returns in the years 1963-1990 (IS). 'ew' is equal-weight, 'vw' is value-weight. We manually group numerators into themes from the literature. Strategies are signed to have positive mean returns IS. 'Pct Short' is the share of strategies that short the ratio. 't-stat' and 'Mean Return' are averages across the 65 possible denominators. 'Mean Return' is in bps per month. 'Mean return OOS/IS' is the mean in either 1991-2004 or 1991-2022 (OOS), divided by the mean IS. Data mining can uncover themes from the literature like investment, external financing, and accruals, decades before they are published. For all themes, predictability persists out-of-sample.

| Numerator (Stock Weight) | 1963-1990 (IS) | | | 1991-2004 | 1991-2022 |
	Pct Short	t-stat	Mean Return	Mean Return OOS / IS	
ΔCapital surplus (ew)	100	5.8	0.67	1.04	0.94
ΔCommon stock (ew)	100	5.8	0.69	0.80	0.55
ΔLiabilities (ew)	100	5.7	0.74	0.87	0.56
ΔCurrent liabilities (ew)	100	5.4	0.60	1.04	0.56
ΔInventories (ew)	100	5.3	0.64	1.46	0.80
ΔDebt in current liab (ew)	100	5.2	0.47	0.30	0.32
Stock issuance (ew)	100	5.2	0.89	1.03	0.80
ΔLong-term debt (ew)	100	5.1	0.53	1.31	0.75
ΔNotes payable st (ew)	100	5.1	0.46	0.16	0.25
ΔInterest expense (ew)	100	5.1	0.58	1.01	0.80
ΔPPE net (ew)	100	4.8	0.73	1.41	0.75
ΔPPE gross (ew)	100	4.7	0.73	1.15	0.61
Retained earnings restatement (ew)	100	4.6	0.54	1.38	0.70
ΔAssets (ew)	100	4.5	0.73	1.63	0.94
Stock repurchases (ew)	0	4.3	0.38	0.27	0.63
ΔConvertible debt and stock (ew)	100	4.1	0.42	1.47	1.19
ΔCapital surplus (vw)	100	4.0	0.57	0.72	0.64
ΔCost of goods sold (ew)	100	3.9	0.49	1.41	0.84
Long-term debt issuance (ew)	89	3.9	0.48	1.29	0.71
ΔInvested capital (ew)	100	3.9	0.63	2.16	1.20

Table A.3: Themes from Mining Accounting Ratios in 2000

Table reports the 20 accounting ratio numerator and stock weight (equal- or value-) combinations with the largest mean t-stats using returns in the years 1963-2000 (IS). 'ew' is equal-weight, 'vw' is value-weight. We manually group numerators into themes from the literature. Strategies are signed to have positive mean returns IS. 'Pct Short' is the share of strategies that short the ratio. 't-stat' and 'Mean Return' are averages across the 65 possible denominators. 'Mean Return' is in bps per month. 'Mean return OOS/IS' is the mean in either 2001-2004 or 2001-2022 (OOS), divided by the mean IS. Data mining can uncover themes from the literature like investment, external financing, and accruals, decades before they are published. For all themes, predictability persists out-of-sample.

| Numerator (Stock Weight) | 1963-2000 (IS) | | | 2001-2004 | 2001-2022 |
	Pct Short	t-stat	Mean Return	Mean Return OOS / IS	
ΔInventories (ew)	100	6.9	0.76	0.72	0.33
ΔLong-term debt (ew)	100	6.4	0.60	0.81	0.38
ΔCommon stock (ew)	100	6.3	0.65	0.83	0.46
ΔPPE net (ew)	100	6.3	0.82	1.10	0.37
ΔCurrent liabilities (ew)	100	6.1	0.61	0.93	0.33
ΔInterest expense (ew)	100	6.1	0.61	0.45	0.58
ΔLiabilities (ew)	100	6.0	0.71	0.87	0.44
ΔPPE gross (ew)	100	5.9	0.78	0.87	0.30
ΔDebt subordinated convertible (ew)	100	5.4	0.71	1.15	0.62
ΔDebt convertible (ew)	100	5.4	0.61	1.72	0.74
Retained earnings restatement (ew)	100	5.4	0.61	1.07	0.29
ΔInvested capital (ew)	100	5.3	0.83	1.55	0.57
Merger sales contrib (ew)	100	5.2	0.53	0.93	0.51
ΔAssets (ew)	100	5.2	0.86	1.33	0.53
ΔCapital surplus (ew)	100	5.2	0.69	0.86	0.85
ΔCapital expenditure (ew)	100	5.2	0.53	1.67	0.64
ΔCost of goods sold (ew)	100	5.1	0.57	0.67	0.40
ΔNum employees (ew)	100	5.0	0.59	1.42	0.52
ΔIntangible assets (ew)	100	5.0	0.49	1.89	0.61
ΔDebt in current liab (ew)	100	4.9	0.40	0.03	0.32

Table A.4: Themes from Mining Accounting Ratios in 2010

Table reports the 20 accounting ratio numerator and stock weight (equal- or value-) combinations with the largest mean t-stats using returns in the years 1963-2010 (IS). 'ew' is equal-weight, 'vw' is value-weight. We manually group numerators into themes from the literature. Strategies are signed to have positive mean returns IS. 'Pct Short' is the share of strategies that short the ratio. 't-stat' and 'Mean Return' are averages across the 65 possible denominators. 'Mean Return' is in bps per month. 'Mean return OOS/IS' is the mean in either 2011-2015 or 2011-2022 (OOS), divided by the mean IS. Data mining can uncover themes from the literature like investment, external financing, and accruals, decades before they are published. For all themes, predictability persists out-of-sample.

| Numerator (Stock Weight) | 1963-2010 (IS) | | | 2011-2014 | 2011-2022 |
	Pct Short	t-stat	Mean Return	Mean Return OOS / IS	
ΔLong-term debt (ew)	100	6.5	0.54	0.64	0.25
ΔInventories (ew)	100	6.5	0.64	0.48	0.47
ΔLiabilities (ew)	100	6.4	0.68	0.52	0.14
ΔCommon stock (ew)	100	6.3	0.59	0.24	0.41
ΔInterest expense (ew)	100	6.3	0.57	0.61	0.51
ΔPPE net (ew)	100	6.1	0.72	0.58	0.37
ΔCurrent liabilities (ew)	100	5.8	0.54	0.17	0.24
ΔDebt convertible (ew)	100	5.7	0.60	0.80	0.60
Merger sales contrib (ew)	100	5.5	0.47	0.16	0.52
ΔAssets (ew)	100	5.5	0.81	0.40	0.35
ΔInvested capital (ew)	100	5.5	0.78	0.43	0.48
ΔIntangible assets (ew)	100	5.4	0.50	0.30	0.23
ΔPPE gross (ew)	100	5.3	0.65	0.52	0.45
ΔConvertible debt and stock (ew)	100	5.0	0.47	1.05	0.89
ΔNum employees (ew)	100	5.0	0.54	0.25	0.53
Retained earnings restatement (ew)	100	5.0	0.51	0.53	0.19
ΔCapital surplus (ew)	100	4.8	0.65	0.54	0.97
ΔDebt subordinated convertible (ew)	100	4.8	0.63	0.34	0.86
ΔDebt in current liab (ew)	100	4.8	0.35	0.32	0.25
ΔCapital expenditure (ew)	100	4.7	0.47	0.36	0.88

A.3 Robustness: Does Risk-Based Theory Help?

Figure A.1: Agnostic and Mispricing Predictors vs Data-Mining

The plot shows long-short returns in event time, where the event is the end of the original sample periods. Predictor returns are normalized to average 100 bps in the original samples. Data-mined predictors come from ratios or scaled first differences from 240 accounting variables (Section 2.1). For all categories of theory, peer review and data mining lead to similar post-sample returns.

(a) Mispricing-Based
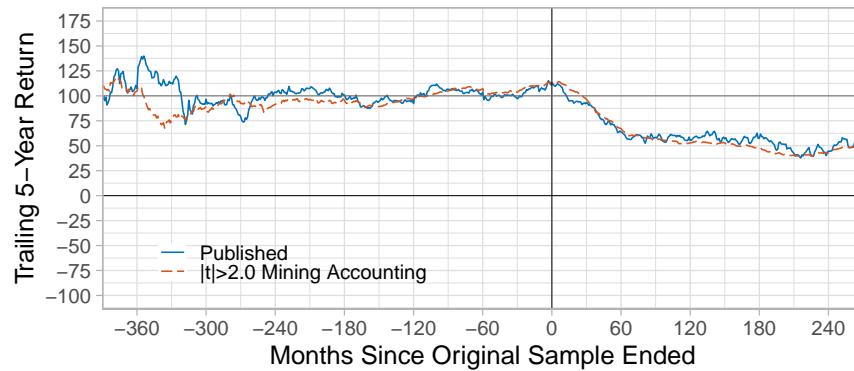
(b) Agnostic
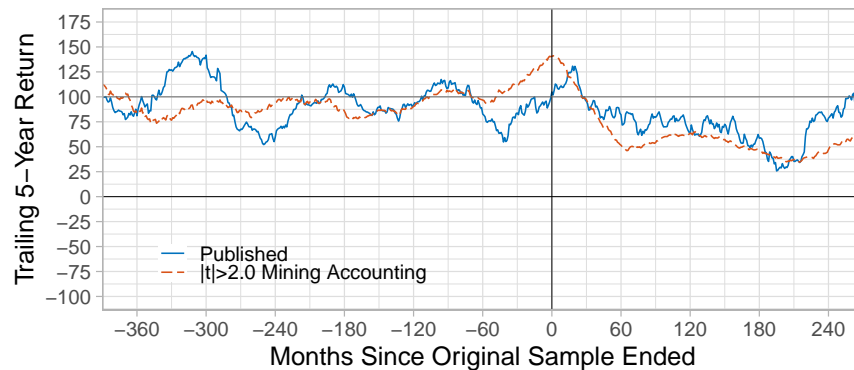
Figure A.2: Published annual accounting predictors against data-mined benchmarks

We compare published strategies to data mined benchmarks based on original-sample t-stats. 'Published' includes only the subset of 84 predictors that are categorized as accounting predictors in the Chen and Zimmermann (2022) dataset and that are based on the annual version of the Compustat database. '$|t| > 2.0$ Mining Accounting' is a data-mined benchmark formed by filtering 29,000 strategies for $|t| > 2.0$ in published predictors' original sample periods. The plot shows long-short returns in event time, where the event is the end of the original sample periods. All predictor returns are normalized to average 100 bps in the original samples. Results are similar to our benchmark results in Figure 1.
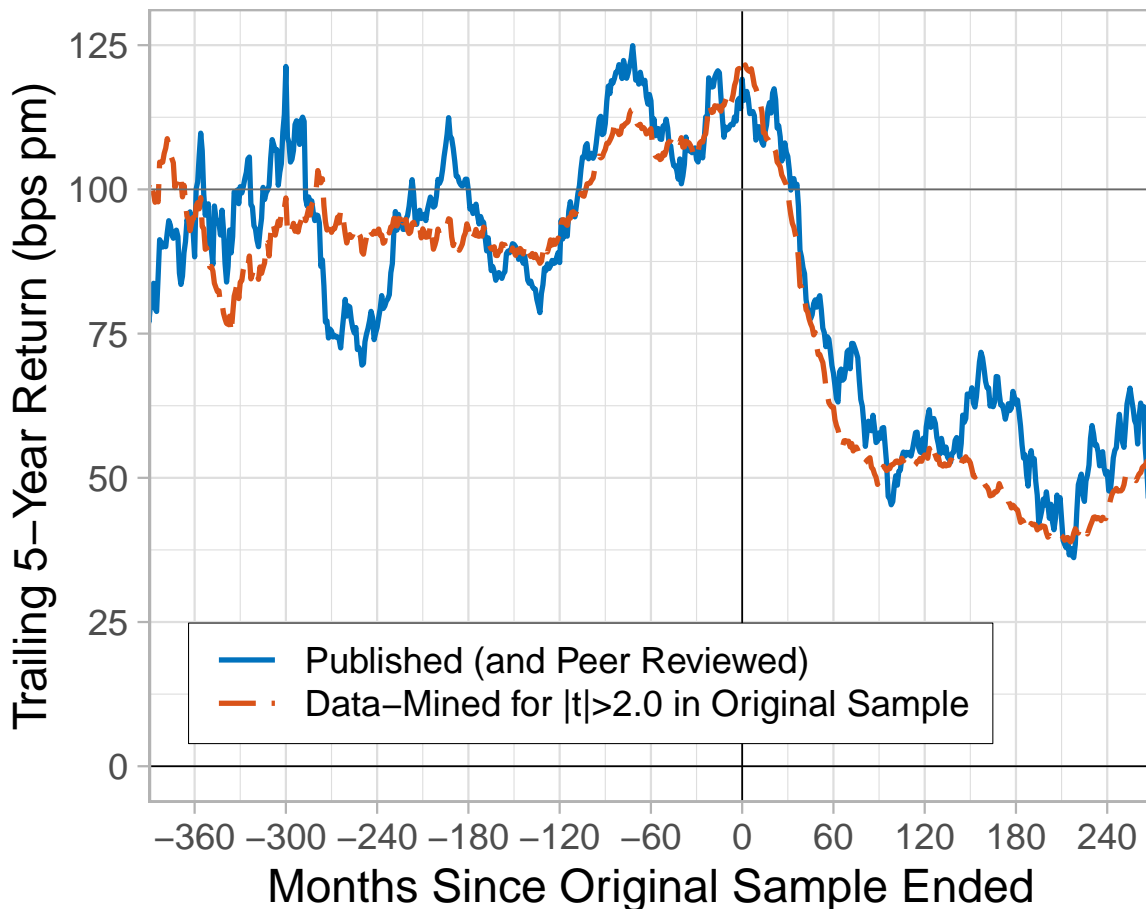
Table A.5: Signals by Theory and Published Journal

This table lists the number of signals by theory and published journal. Finance journals find risk explanations more frequently than accounting journals, but risk explanations still account for a small minority of predictors in finance journals.

	Agnostic	Mispricing	Risk
AR	1	14	0
BAR	0	1	0
Book	2	0	0
CAR	0	1	0
FAJ	1	1	0
JAE	3	14	0
JAR	3	2	0
JBFA	0	1	0
JEmpFin	0	1	0
JF	16	35	12
JFE	16	22	6
JFM	0	2	0
JFQA	0	3	2
JFR	0	0	1
JOIM	0	1	0
JPE	0	0	3
JPM	1	0	0
MS	0	2	2
Other	0	1	0
RAS	0	5	1
RED	0	0	1
RFQA	0	1	0
RFS	0	7	7
ROF	0	1	3
WP	1	1	0

Figure A.3: Abnormal CAPM Returns

The plot shows the abnormal return of long-short returns of published predictors in event time, where the event is the end of the original sample periods. We calculate abnormal returns as $abnormal_{i,t} = r_{i,t} - \beta_{i,t} r_t^e$. We calculate beta separately for the original sample period, and after the original sample period. We keep the abnormal returns if the t-statistic is greater than one during the original sample period. Each abnormal return is normalized so that its mean original-sample is 100 bps per month. Predictors are grouped by theory category based on the arguments in the original papers (Table 4). We average abnormal returns across predictors within each month and then take the trailing 5-year average for readability. For all categories of theory, predictability decays by roughly 50% post-sample. If anything, risk-based predictors decay more than other predictors.
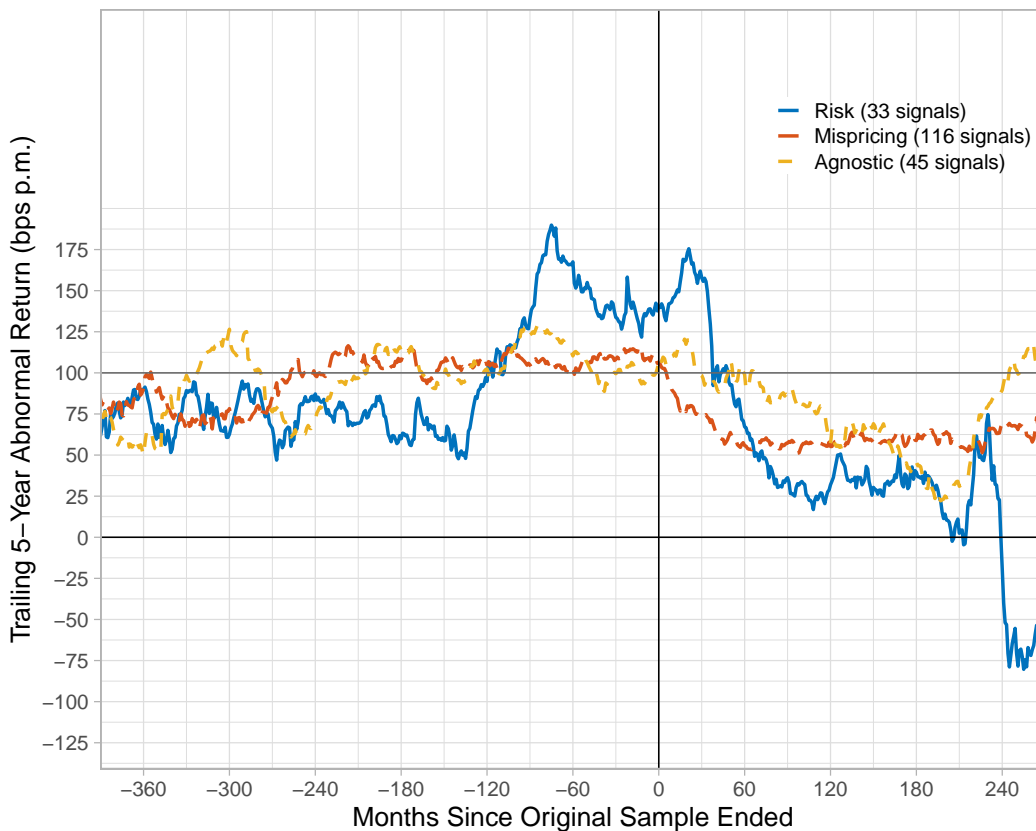
Figure A.4: Controlling for Sample Mean Returns, t-stats, Correlations

We compare published strategies to data mined benchmarks based on original-sample t-stats (as in Figures 1 and 4) but now we drop data-mined strategies if they have t-stats that differ by more than 10% or mean returns that differ by more than 10% (short-dash). We additionally drop data-mined strategies that are more than 10% correlated with published strategies in the original sample (long-dash). The main results are robust to an even finer control for in-sample mean returns compared to Figure 6. The plot omits 21 strategies without matched strategies after the filtering: "AccrualsBM," "AnalystRevision," "AssetGrowth," "BM," "BMdec," "Beta," "DivYieldST," "FEPS," "Frontier," "Mom6mJunk," "MomRev," "MomSeasonShort," "MomVol," "NOA," "Price," "Recomm_ShortInterest," "ResidualMomentum," "dCPVolSpread," "dNoa," "retConglomerate," and "roaq."

(a) All

(b) Risk-based

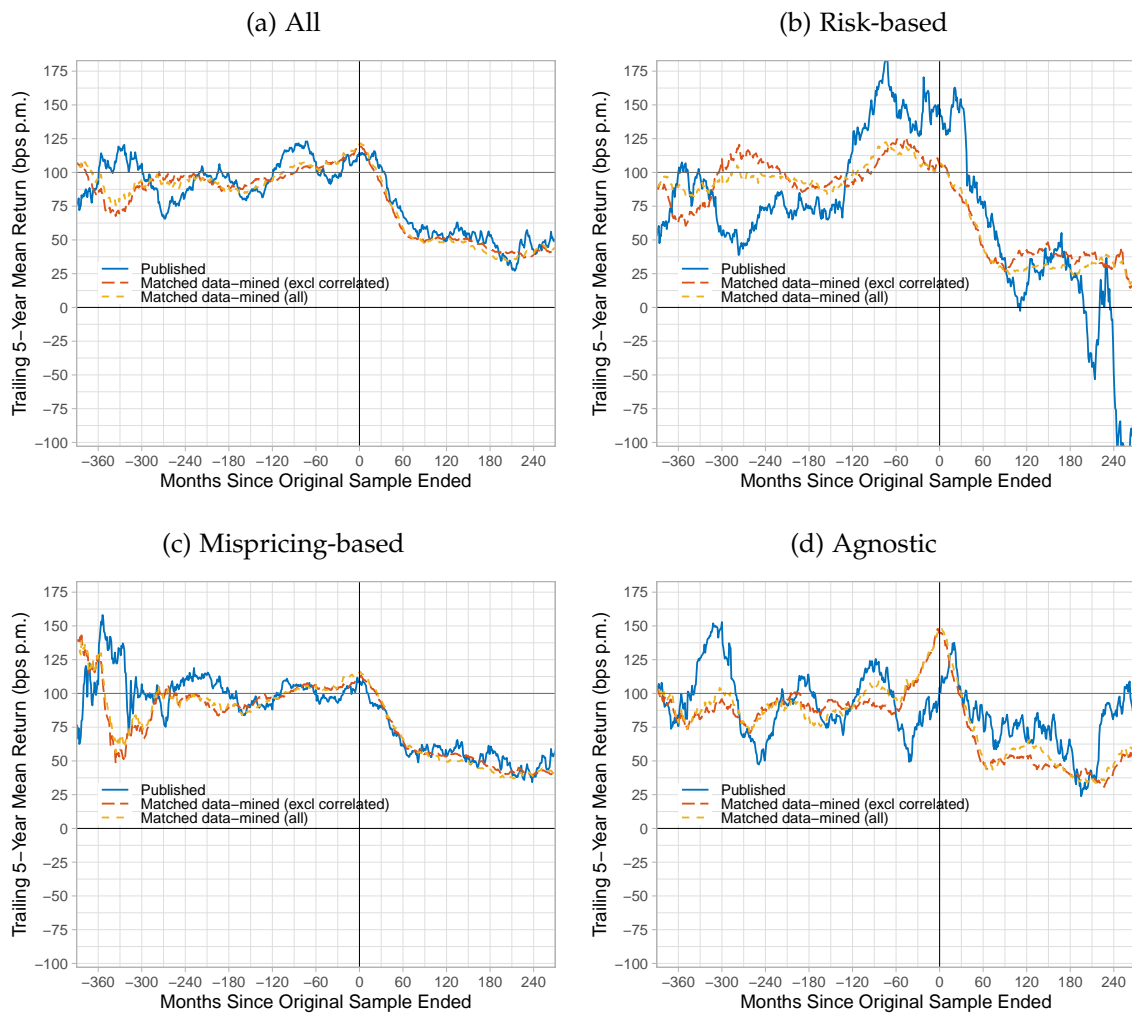(c) Mispricing-based

(d) Agnostic

44

Table A.6: Regression Estimates of Risk vs Mispricing Effects on Predictability Decay

We regress monthly long-short returns on indicator variables to quantify the effects of peer-reviewed risk vs mispricing explanations on predictability decay. "Post-Sample" is 1 if the month occurs after the predictor's sample ends and is zero otherwise. "Post-Pub" is defined similarly. "Risk" is 1 if peer review argues for a risk-based explanation (Table 4) and 0 otherwise. "Mispricing" and "Post-2004" are defined similarly. Parentheses show standard errors clustered by month. "Null: Risk No Decay" shows the p-value that tests whether risk-based returns do not decrease post-sample ((1) and (3)) or post-publication ((2) and (4)). Risk-based predictors decay more than other predictors, but the difference is only marginally significant. The decay in risk-based predictors overall is highly significant.

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	68.2	68.2	68.2	68.2	69.6
	(3.9)	(3.9)	(3.9)	(3.9)	(4.1)
Post-Sample	-26.9	-21.8	-25.8	-22.0	-2.7
	(5.5)	(7.3)	(6.1)	(9.7)	(9.1)
Post-Pub		-6.3		-4.6	
		(7.9)		(11.4)	
Post-Sample x Risk	-19.4	-10.0	-20.5	-9.8	-16.4
	(7.4)	(9.8)	(7.9)	(11.2)	(7.3)
Post-Pub x Risk		-12.4		-14.2	
		(12.9)		(14.2)	
Post-Sample x Mispricing			-1.6	0.3	
			(5.1)	(8.8)	
Post-Pub x Mispricing				-2.4	
				(10.0)	
Post-2004					-33.6
					(10.9)
Null: Risk No Decay	< 0.1%	< 0.1%	< 0.1%	< 0.1%	< 0.1%

A.4 Proof of Proposition 1

Proof. Let $s_i \equiv \text{Sign}(\bar{r}_i)$. Then rewrite the expected decay conditioning on $|\bar{r}_i|$ and i belonging to some set $\mathcal{S}$

$$E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{S}\right] = E\left[s_i\left(\bar{r}_i - \bar{r}_i^{OOS}\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{7}$$

$$= E\left[s_i\left(\bar{r}_i - \mu_i - \Delta\mu_i - \bar{\varepsilon}_i^{OOS}\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{8}$$

$$= E\left[s_i\left(\bar{r}_i - \mu_i - \Delta\mu_i\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{9}$$

$$= E\left[s_i\bar{r}_i - s_i\mu_i - s_i\Delta\mu_i ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{10}$$

where the first line uses Equation (1) and the second line uses Equation (3).

Then plug in $\mathcal{S} = \mathcal{T}$ and then use the two assumptions of the proposition (Equations (5) and (6))

$$E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{T}\right] = E\left[s_i\bar{r}_i - s_i\mu_i + s_i\Delta\mu_i ||\bar{r}_i|, i \in \mathcal{T}\right]$$

$$< E\left[s_i\bar{r}_i - s_i\mu_i + s_i\Delta\mu_i ||\bar{r}_i|, i \in \mathcal{D}\right]$$

$$= E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{D}\right]$$

Integrating over $|\bar{r}_i| > h$ finishes the proof. □

A.5   Additional Robustness

Table A.7: 20 Data-Mined Predictors That Perform Similarly to Banz's Size (1981)

Table lists 20 of the 221 data-mined signals that performed similarly to Banz's (1981) size in the original sample period. Signals are ranked according to the absolute difference in mean original-sample return. Sign = -1 indicates that a high signal implies a lower mean return original-sample. Data mining leads to similar out-of-sample performance.

Similarity Rank	Signal	Sign	Mean Return (% Monthly)	
			1926-1975	1976-2021
Peer-Reviewed				
	Size (Banz 1981)	-1	0.50	0.19
Data-Mined				
1	Δ[Equity liq value]/lag[Sales]	-1	0.50	0.77
2	Δ[Invest tax credit inc ac]/lag[Pref stock redemp val]	-1	0.49	-0.13
3	[Cost of goods sold]/[Capex PPE sch V]	1	0.50	0.82
4	Δ[Assets]/lag[Preferred stock liquidat]	-1	0.49	0.20
5	Δ[Equity liq value]/lag[Curr assets other sundry]	-1	0.48	0.73
6	Δ[Equity liq value]/lag[Current liabilities]	-1	0.48	0.85
7	Δ[Receivables]/lag[Pref stock redemp val]	-1	0.48	0.19
8	[Market equity FYE]/[Invested capital]	-1	0.49	0.73
9	Δ[Current assets]/lag[Invest tax credit inc ac]	-1	0.52	0.33
10	Δ[Assets]/lag[Pref stock redemp val]	-1	0.47	0.25
	. . .			
101	Δ[Sales]/lag[Current liabilities]	-1	0.40	0.77
102	Δ[Interest expense]/lag[Cost of goods sold]	-1	0.40	0.71
103	Δ[Interest expense]/lag[Num employees]	-1	0.40	0.68
104	Δ[Operating expenses]/lag[Long-term debt]	-1	0.40	0.48
105	Δ[Operating expenses]/lag[Current liabilities]	-1	0.40	0.90
	. . .			
234	[Gross profit]/[Earnings before interest]	1	0.35	0.19
235	[Market equity FYE]/[Capex PPE sch V]	-1	0.35	0.30
236	[PPE land and improvement]/[Pension retirem expense]	-1	0.64	-0.00
237	[Interest expense]/[Cost of goods sold]	-1	0.35	0.63
238	[Operating expenses]/[Op income after deprec]	1	0.35	0.15
	Mean Data-Mined		0.43	0.44

A.5.1   Numbers of Correlated Data-Mined Strategies

Figure A.5 and A.6 plot the number of total data-mined strategies available at a time, the number of spanned data-mined portfolios, and the number of unspanned data-mined portfolios for both procedures. Figure A.5 displays the result when spanned means

47

correlated with any academic signal, and Figure A.6 shows the result when spanned means correlated with a linear combination of the 5-factor of the academic signals. In the first case, even by the end of the sample, the number of unspanned portfolios remains close to 20,000. In the second case, the number of unspanned portfolios is 'only' around 5000 by the end of the sample.

Figure A.5: Number of Unspanned and Spanned Data-Mined Strategies: Individual Correlation

The figure shows the number of spanned and unspanned strategies by each date. Unspanned means that the correlation of the matched data-mined strategy with all published strategies by the time it is matched is lower than 50%, and spanned means it is higher.
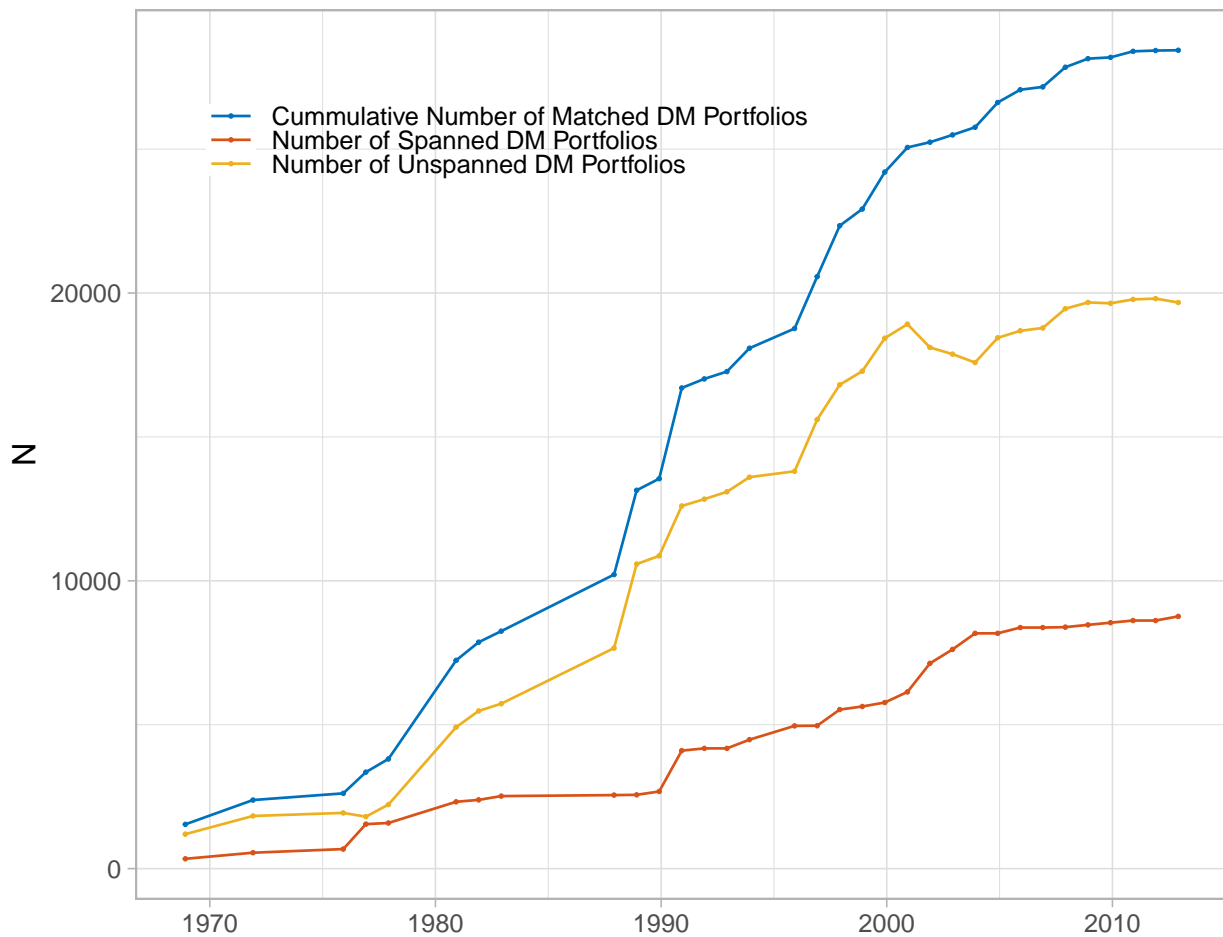
Figure A.6: Number of Unspanned and Spanned Data-Mined Strategies: PPCA

The figure shows the number of spanned and unspanned strategies by each date. Spanned means that the adjusted R2 of a regression of the data mined strategy against the first 5 principal components of the academic signals available at the time is more than 0.25. We use probabilistic PCA to deal with the incomplete panel of academic signals and require a minimum of 30 observations to run the regression.
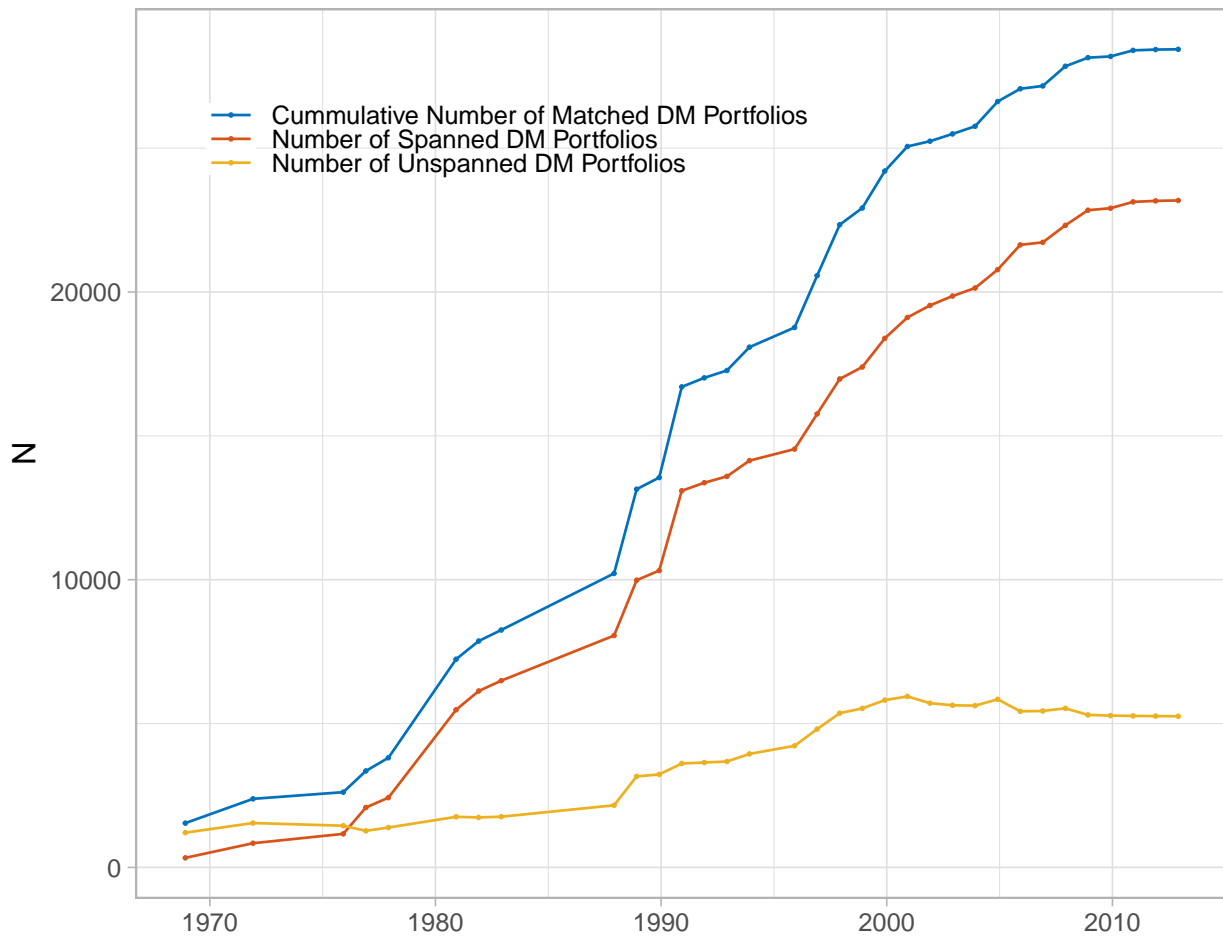
References

Abarbanell, Jeffery S and Brian J Bushee (1998). "Abnormal returns to a fundamental analysis strategy". In: *Accounting Review*, pp. 19–45.

Akerlof, George A and Pascal Michaillat (2018). "Persistence of false paradigms in low-power sciences". In: *Proceedings of the National Academy of Sciences* 115.52, pp. 13228–13233.

Amihud, Yakov (2002). "Illiquidity and stock returns: cross-section and time-series effects". In: *Journal of financial markets* 5.1, pp. 31–56.

Ankel-Peters, Jörg, Nathan Fiala, and Florian Neubauer (2023). "Is Economics Self-Correcting? Replications in the American Economic Review". In.

Asness, Clifford S, Tobias J Moskowitz, and Lasse Heje Pedersen (2013). "Value and momentum everywhere". In: *The journal of finance* 68.3, pp. 929–985.

Bali, Turan G, Heiner Beckmeyer, and Timo Wiedemann (2023). "Expected Mispricing". In: *Available at SSRN*.

Bali, Turan G, Robert F Engle, and Scott Murray (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Banz, Rolf W (1981). "The relationship between return and market value of common stocks". In: *Journal of financial economics* 9.1, pp. 3–18.

Barberis, Nicholas (2018). "Psychology-based models of asset prices and trading volume". In: *Handbook of behavioral economics: applications and foundations 1*. Vol. 1. Elsevier, pp. 79–175.

Barry, Christopher B and Stephen J Brown (1984). "Differential information and the small firm effect". In: *Journal of financial economics* 13.2, pp. 283–294.

Bender, Svetlana et al. (2022). "Millionaires speak: What drives their personal investment decisions?" In: *Journal of Financial Economics* 146.1, pp. 305–330.

Bessembinder, Hendrik, Aaron Burt, and Christopher M Hrdlicka (2023). "Time Series Variation in the Factor Zoo". In: *Aaron Paul and Hrdlicka, Christopher M., Time Series Variation in the Factor Zoo*.

Böll, Julian et al. (2024). "Following the Footprints: Towards a Taxonomy of the Factor Zoo". In: *Available at SSRN*.

Boudoukh, Jacob et al. (2007). "On the importance of measuring payout yield: Implications for empirical asset pricing". In: *The Journal of Finance* 62.2, pp. 877–915.

Brav, Alon and John B Heaton (2002). "Competing theories of financial anomalies". In: *The Review of Financial Studies* 15.2, pp. 575–606.

Campbell, John Y and Tuomo Vuolteenaho (2004). "Bad beta, good beta". In: *American Economic Review* 94.5, pp. 1249–1275.

Celerier, Claire, Boris Vallee, and Alexey Vasilenko (2022). "What Drives Finance professors' Wages?" In.

Chen, Andrew Y (2018). "A general equilibrium model of the value premium with time-varying risk premia". In: *The Review of Asset Pricing Studies* 8.2, pp. 337–374.

— (2021). "The Limits of p-Hacking: Some Thought Experiments". In: *The Journal of Finance* 76.5, pp. 2447–2480.

— (2024). "Most claimed statistical findings in cross-sectional return predictability are likely true". In: *arXiv preprint arXiv:2206.15365*.

Chen, Andrew Y and Chukwuma Dim (2023). "High-Throughput Asset Pricing". In: *arXiv preprint arXiv:2311.10685*.

Chen, Andrew Y and Jack McCoy (2024). "Missing values handling for machine learning portfolios". In: *Journal of Financial Economics* 155, p. 103815.

Chen, Andrew Y and Mihail Velikov (2022). "Zeroing in on the Expected Returns of Anomalies". In: *Journal of Financial and Quantitative Analysis*.

Chen, Andrew Y and Tom Zimmermann (2020). "Publication bias and the cross-section of stock returns". In: *The Review of Asset Pricing Studies* 10.2, pp. 249–289.

— (2023). "Publication Bias in Asset Pricing Research". In: *Oxford Research Encyclopedia of Economics and Finance*.

— (2022). "Open Source Cross Sectional Asset Pricing". In: *Critical Finance Review*.

Chinco, Alex, Samuel M Hartzmark, and Abigail B Sussman (2022). "A new test of risk factor relevance". In: *The Journal of Finance* 77.4, pp. 2183–2238.

Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). "Anomalies and false rejections". In: *The Review of Financial Studies* 33.5, pp. 2134–2179.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong (2014). "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?" In: *Journal of Accounting and Economics* 58.1, pp. 41–58.

Cochrane, John H (2009). *Asset pricing: Revised edition*. Princeton university press.

— (2017). "Macro-finance". In: *Review of Finance* 21.3, pp. 945–985.

Da, Zhi, Umit G Gurun, and Mitch Warachka (2014). "Frog in the pan: Continuous information and momentum". In: *The review of financial studies* 27.7, pp. 2171–2218.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam (1998). "Investor psychology and security market under-and overreactions". In: *the Journal of Finance* 53.6, pp. 1839–1885.

DeMiguel, Victor et al. (2020). "A Transaction-Cost Perspective on the Multitude of Firm Characteristics". In: *The Review of Financial Studies* 33.5, pp. 2180–2222.

Doran, James and Colbrin Wright (2007). "What Really Matters When Buying and Selling Stocks?" In: *Financial Education* 8.1, pp. 35–61.

Fama, Eugene F (1970). "Efficient capital markets: A review of theory and empirical work". In: *The journal of Finance* 25.2, pp. 383–417.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". In: *the Journal of Finance* 47.2, pp. 427–465.

— (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56.

— (2015). "A five-factor asset pricing model". In: *Journal of financial economics* 116.1, pp. 1–22.

— (2018). "Choosing factors". In: *Journal of financial economics* 128.2, pp. 234–252.

Foster, George, Chris Olsen, and Terry Shevlin (1984). "Earnings releases, anomalies, and the behavior of security returns". In: *Accounting Review*, pp. 574–603.

Frazzini, Andrea and Lasse Heje Pedersen (2014). "Betting against beta". In: *Journal of Financial Economics* 111.1, pp. 1–25.

Frey, Jonas (2023). "Which stock return predictors reflect mispricing?" In: *Available at SSRN*.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber (2020). "Dissecting characteristics nonparametrically". In: *The Review of Financial Studies* 33.5, pp. 2326–2377.

Gabaix, Xavier (2008). "Variable rare disasters: A tractable theory of ten puzzles in macro-finance". In: *American Economic Review* 98.2, pp. 64–67.

Gomes, Joao, Leonid Kogan, and Lu Zhang (2003). "Equilibrium cross section of returns". In: *Journal of Political Economy* 111.4, pp. 693–732.

Goto, Shingo and Toru Yamada (2022). "False Alpha and Missed Alpha: An Out-of-Sample Mining Expedition". In: *Working Paper*.

Green, Jeremiah, John RM Hand, and X Frank Zhang (2017). "The characteristics that provide independent information about average US monthly stock returns". In: *The Review of Financial Studies* 30.12, pp. 4389–4436.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Han, Yufeng et al. (2022). "Expected Stock Returns and Firm Characteristics: E-ENet, Assessment, and Implications". In: *Working Paper*.

Harvey, Campbell R (2017). "Presidential address: The scientific outlook in financial economics". In: *The Journal of Finance* 72.4, pp. 1399–1440.

Harvey, Campbell R and Yan Liu (2020). "False (and missed) discoveries in financial economics". In: *The Journal of Finance* 75.5, pp. 2503–2553.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Hasler, Mathias (2023). "Looking under the hood of data-mining". In: *Available at SSRN 4279944*.

Haugen, Robert A and Nardin L Baker (1996). "Commonality in the determinants of expected stock returns". In: *Journal of financial economics* 41.3, pp. 401–439.

Heston, Steven L and Ronnie Sadka (2008). "Seasonality in the cross-section of stock returns". In: *Journal of Financial Economics* 87.2, pp. 418–445.

Holcblat, Benjamin, Abraham Lioui, and Michael Weber (2022). "Anomaly or possible risk factor? Simple-to-use tests". In: *Simple-To-Use Tests (April 3, 2022)*.

Holden, Craig W and Avanidhar Subrahmanyam (2002). "News events, information acquisition, and serial correlation". In: *The Journal of Business* 75.1, pp. 1–32.

Hong, Harrison and Jeremy C Stein (1999). "A unified theory of underreaction, momentum trading, and overreaction in asset markets". In: *The Journal of finance* 54.6, pp. 2143–2184.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency". In: *The Journal of finance* 48.1, pp. 65–91.

Jensen, Michael C. and George A. Benington (1970). "Random Walks and Technical Theories: Some Additional Evidence". In: *The Journal of Finance* 25.2, pp. 469–482.

Jensen, Theis Ingerslev, Bryan T Kelly, et al. (2022). "Machine learning and the implementable efficient frontier". In: *Swiss Finance Institute Research Paper* 22-63.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Kelly, Bryan, Ronen Israel, and Tobias Moskowitz (2020). "Can Machines "Learn" Finance?" In: *Journal of Investment Management* 18.2.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020). "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2, pp. 271–292.

Krusell, Per and Anthony A Smith Jr (1998). "Income and wealth heterogeneity in the macroeconomy". In: *Journal of political Economy* 106.5, pp. 867–896.

Lettau, Martin and Jessica A Wachter (2007). "Why is long-horizon equity less risky? A duration-based explanation of the value premium". In: *The journal of finance* 62.1, pp. 55–92.

Lo, Andrew W and A Craig MacKinlay (1990). "Data-snooping biases in tests of financial asset pricing models". In: *The Review of Financial Studies* 3.3, pp. 431–467.

Lopez-Lira, Alejandro and Nikolai L Roussanov (2020). "Do Common Factors Really Explain the Cross-Section of Stock Returns?" In: *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Marcet, Albert (1991). "Solving non-linear stochastic models by parameterizing expectations: An application to asset pricing with production". In.

Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R Green (1995). *Microeconomic theory*. Vol. 1. Oxford university press New York.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" In: *The Journal of Finance* 71.1, pp. 5–32.

Moritz, Benjamin and Tom Zimmermann (2016). "Tree-based conditional portfolio sorts: The relation between past and future stock returns". In: *Available at SSRN 2740751*.

Moskowitz, Tobias J and Mark Grinblatt (1999). "Do industries explain momentum?" In: *The Journal of finance* 54.4, pp. 1249–1290.

Mukhlynina, Liliya and Kjell G Nyborg (2020). "The Choice of Valuation Techniques in Practice: Education Versus Profession". In: *Critical Finance Review* 9.1-2, pp. 201–265.

Papanikolaou, Dimitris (2011). "Investment shocks and asset prices". In: *Journal of Political Economy* 119.4, pp. 639–685.

Pástor, Ľuboš and Robert F Stambaugh (2003). "Liquidity risk and expected stock returns". In: *Journal of Political economy* 111.3, pp. 642–685.

Pontiff, Jeffrey and Artemiza Woodgate (2008). "Share issuance and cross-sectional returns". In: *The Journal of Finance* 63.2, pp. 921–945.

Roweis, Sam (1997). "EM algorithms for PCA and SPCA". In: *Advances in neural information processing systems* 10.

Shiller, Robert J (2003). "From efficient markets theory to behavioral finance". In: *Journal of economic perspectives* 17.1, pp. 83–104.

Sloan, Richard G (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings?" In: *Accounting review*, pp. 289–315.

Soebhag, Amar, Bart Van Vliet, and Patrick Verwijmeren (2024). "Non-standard errors in asset pricing: Mind your sorts". In: *Journal of Empirical Finance*, p. 101517.

Soliman, Mark T (2008). "The use of DuPont analysis by market participants". In: *The Accounting Review* 83.3, pp. 823–853.

Spiess, D Katherine and John Affleck-Graves (1999). "The long-run performance of stock returns following debt offerings". In: *Journal of Financial Economics* 54.1, pp. 45–73.

Stambaugh, Robert F, Jianfeng Yu, and Yu Yuan (2012). "The short of it: Investor sentiment and anomalies". In: *Journal of financial economics* 104.2, pp. 288–302.

Stattman, Dennis (1980). "Book values and stock returns". In: *The Chicago MBA: A journal of selected papers* 4.1, pp. 25–45.

Subrahmanyam, Avanidhar (2018). "Equity market momentum: A synthesis of the literature and suggestions for future work". In: *Pacific-Basin Finance Journal* 51, pp. 291–296.

Sullivan, Ryan, Allan Timmermann, and Halbert White (1999). "Data-snooping, technical trading rule performance, and the bootstrap". In: *The journal of Finance* 54.5, pp. 1647–1691.

— (2001). "Dangers of data mining: The case of calendar effects in stock returns". In: *Journal of Econometrics* 105.1, pp. 249–286.

Sutton, Richard (2019). "The bitter lesson". In: *Incomplete Ideas (blog)* 13.1, p. 38.

Thomas, Jacob K and Huai Zhang (2002). "Inventory changes and future returns". In: *Review of Accounting Studies* 7.2, pp. 163–187.

Titman, Sheridan, KC John Wei, and Feixue Xie (2004). "Capital investments and stock returns". In: *Journal of financial and Quantitative Analysis* 39.4, pp. 677–700.

Tuzel, Selale (2010). "Corporate real estate holdings and the cross-section of stock returns". In: *The Review of Financial Studies* 23.6, pp. 2268–2302.

Walter, Dominik, Rüdiger Weber, and Patrick Weiss (2023). "Non-standard errors in portfolio sorts". In: *Available at SSRN 4164117*.

Yan, Xuemin Sterling and Lingling Zheng (2017). "Fundamental analysis and the cross-section of stock returns: A data-mining approach". In: *The Review of Financial Studies* 30.4, pp. 1382–1423.

Zaffaroni, Paolo and Guofu Zhou (2022). "Asset Pricing: Cross-section Predictability". In: *Available at SSRN 4111428*.

Zhang, Lu (2005). "The value premium". In: *The Journal of Finance* 60.1, pp. 67–103.

Zhao, Wayne Xin et al. (2023). *A Survey of Large Language Models*. arXiv: `2303.18223` `[cs.CL]`.

Zhu, Min (2023). "Evaluating the Efficacy of Multiple Testing Adjustments in Empirical Asset Pricing". In: *Available at SSRN 4396035*.