



Decoding Lazy Prices:

An Analysis of 10-K Filings and Stock Returns

Akanksha, Henry, Hannah, Marti

Meet the team



Akanksha Gavade

Junior
IBE Finance & Industrial
and Systems Engineering



Hannah Gordon

Junior
IBE Financial Engineering



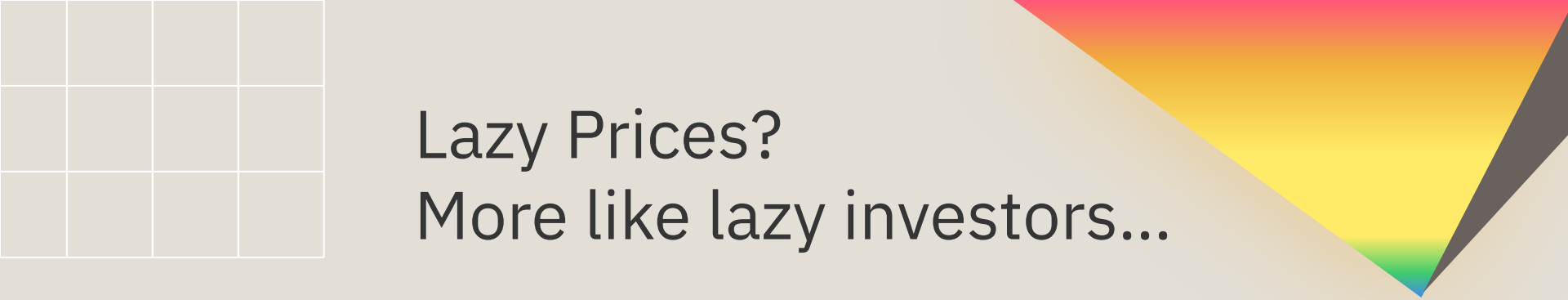
Jose Marti Figueres

Senior
IBE Finance & Chemical
Engineering



Henry Piotrowski

Senior
Finance



Lazy Prices? More like lazy investors...

“

Changes to the language and construction of financial reports also have strong implications for firms' future returns: a portfolio that shorts “changers” and buys “non-changers” earns up to 188 basis points in monthly alphas (over 22% per year) in the future. ”

Are there subtle signals in corporate disclosures that the market initially under reacts to?

Lazy Prices*

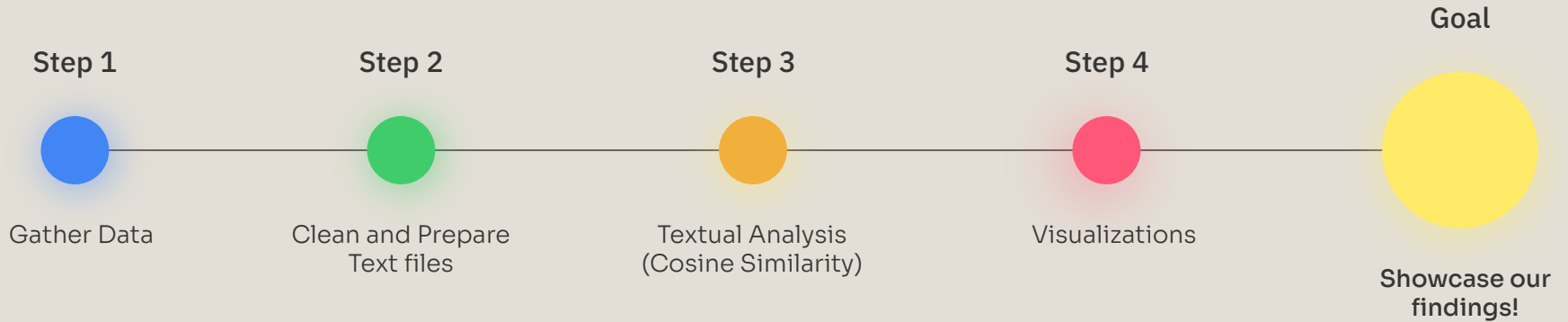
Lauren Cohen
Harvard Business School and NBER

Christopher Malloy
Harvard Business School and NBER

Quoc Nguyen
DePaul University

This Draft: March 7, 2019

Presentation Roadmap



Different ways of gathering the 10Ks from 1993-2024...

Source: <https://sraf.nd.edu/sec-edgar-data/>

1

Use the SEC
EDGAR
downloader

2

Download 50 GB
(zipped) files of all
10-X files

3

15GB 10-K
Document
Dictionaries file,
which contains
word counts for
each word in the
LM dictionary

Coding roadmap

Step 1

Download Text
Data from 10-K
Filings

Step 2

Compute Cosine
Similarity Between
Sequential Filings

Step 3

Create a
Structured
Dataset

Step 4

Merge Filing Data
with Monthly
Stock Returns

Step 5

Separate Cosine
Distances into 5
Bins by Year

Step 6

Export the Final
Cleaned Dataset
for Analysis

Raw Data Format

- 1288776, 20040816, 0001193125-04-141838,20040630,10-K, Google Inc.|101:42,126:16,184:17,186:10,213:1,248:1, ...
- The first six fields before the pipe are: (1) cik, (2) filing_date, (3) accession_number, (4) conformed_period_report, (5) form_type, and (6) company_name.
- The subsequent fields contain the word sequence number in the master dictionary followed by the count for that word. The first field (101:42), indicates that the word "ability "occurred 42 times in the filing.

Cosine Similarity

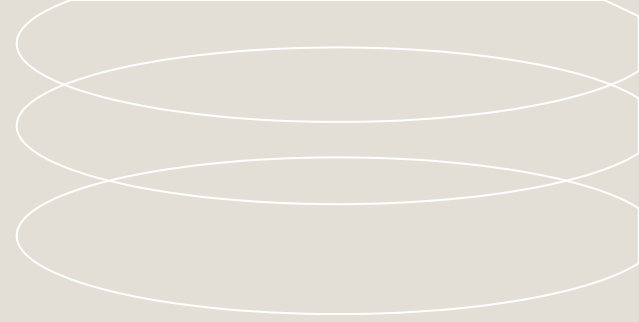
- A **measure of similarity** between two vectors
- **Ignores magnitude** and focuses on **direction**
- **Cosine similarity = 1.0** → No change (identical filings)
- **Closer to 0** → Language is very different

Word Seq	101	126	184	700	...
2022	42	16	17	9	...
2023	50	18	19	11	...

Final Data Format

Symbol	CIK	Year	Filing Date	Cosine Distance	Return Date	Return	Return +1	Cosine Similarity	Bins
MMM	66740	2004	2005-02-24	0.004748691738027722	2005-02-28	0.0	4.9346	0.9952513082619723	5
MMM	66740	2005	2006-02-21	0.010076684725560003	2006-02-28	1.7869	2.0848	0.98992331527444	3
MMM	66740	2006	2007-02-26	0.004970343684424838	2007-02-28	0.3499	2.8537	0.9950296563155752	4
MMM	66740	2007	2008-02-15	0.006057336235305244	2008-02-29	-0.9416	3.1722	0.9939426637646948	3
MMM	66740	2008	2009-02-17	0.007101716524777979	2009-02-27	-14.538	0.9566	0.992898283475222	4
MMM	66740	2009	2010-02-16	0.004239306418998279	2010-02-26	0.2298	9.3709	0.9957606935810017	4
MMM	66740	2010	2011-02-16	0.008228461614766869	2011-02-28	5.5278	4.267	0.9917715383852331	3
MMM	66740	2011	2012-02-16	0.0006635210453727058	2012-02-29	1.7068	1.377	0.9993364789546273	5
MMM	66740	2012	2013-02-14	0.0007362926826388616	2013-02-28	4.0627	1.8379	0.9992637073173611	5
MMM	66740	2013	2014-02-13	0.0007634641683943455	2014-02-28	5.7688	2.2212	0.9992365358316057	5
MMM	66740	2014	2015-02-12	0.0006358092685101457	2015-02-27	4.5441	0.6903	0.9993641907314899	5
MMM	66740	2015	2016-02-11	0.015201703332694105	2016-02-29	4.6225	-2.1939	0.9847982966673059	2
MMM	66740	2016	2017-02-09	0.012178514807067442	2017-02-28	7.2675	6.2217	0.9878214851929326	2
MMM	66740	2017	2018-02-08	0.013599010681544454	2018-02-28	-5.4411	2.6724	0.9864009893184555	2
MMM	66740	2018	2019-02-07	0.001468682193085824	2019-02-28	4.2586	-6.7895	0.9985313178069142	4

Repo Demo



<https://github.com/martifiguere/Final-Project-HPST/tree/main>



Challenges

1

Long running time

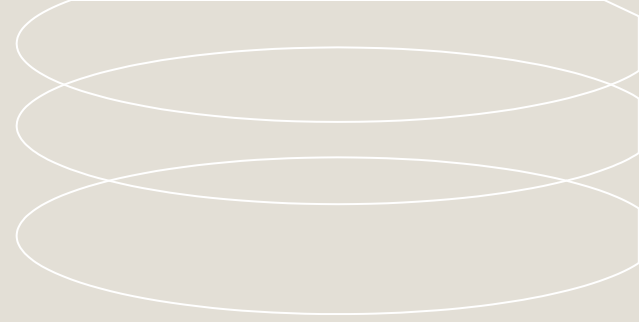
2

Missing Data

3

Intuitive ways to
visualize our data

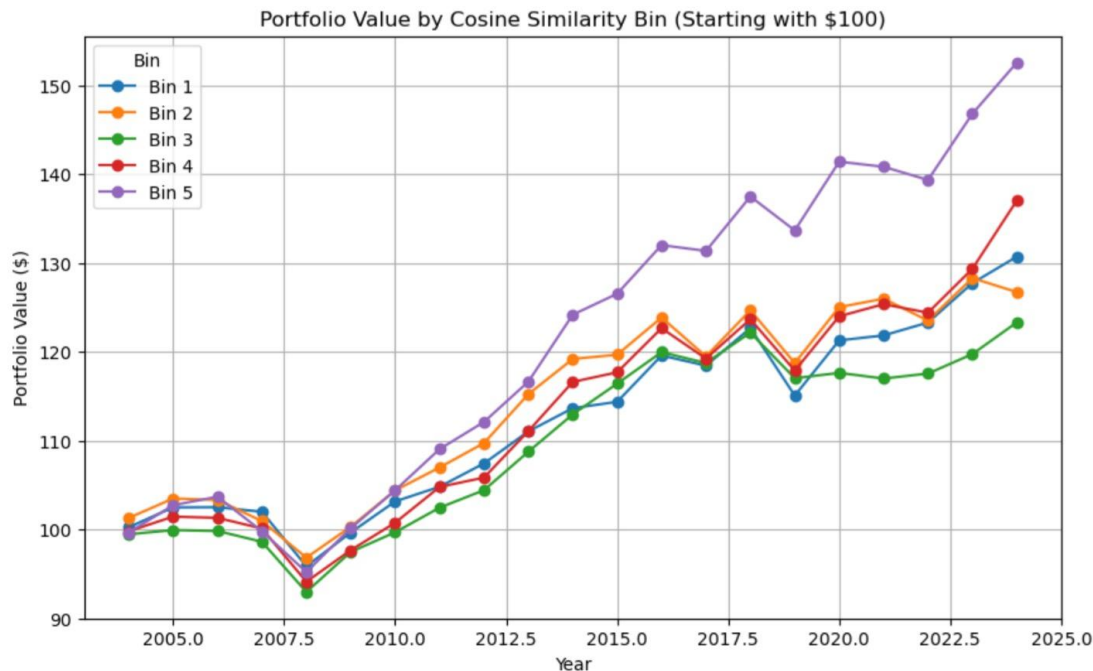
Website Demo



<https://fin377lazyprices.streamlit.app/>

Insights

The line chart shows how \$100 invested in each cosine similarity quintile (Bin 1 = least similar; Bin 5 = most similar) would have grown over time:



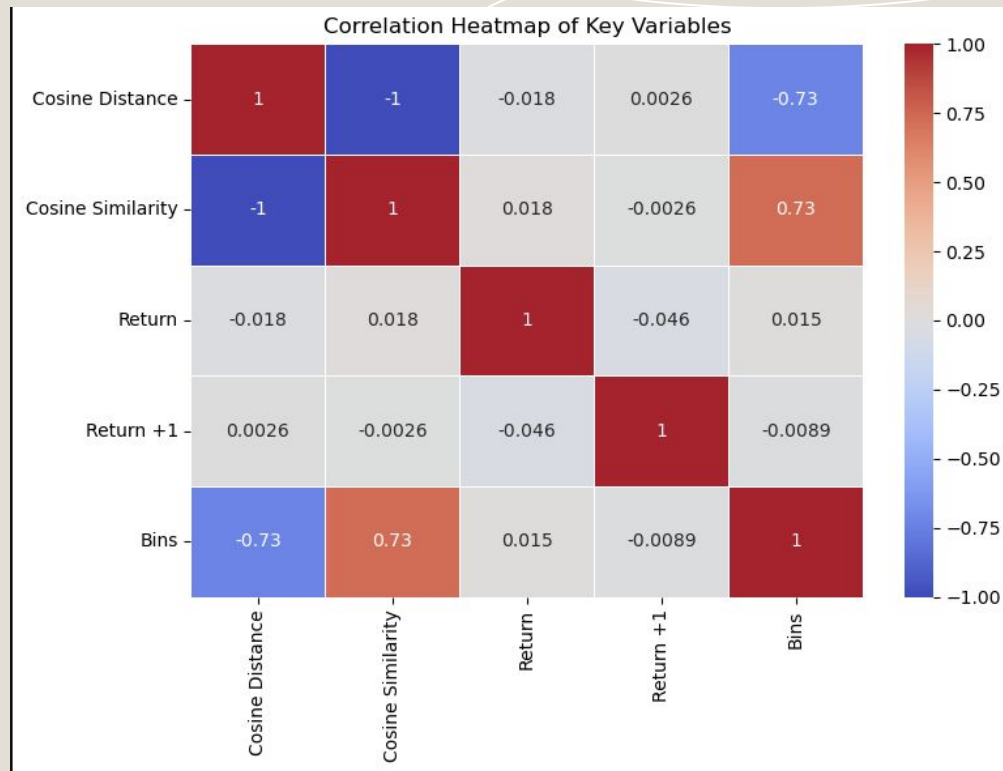
Bin 5 portfolios **outperformed** over the long run.

Bin 2 and 4 had **strong performance, but were outperformed by Bin 5**, peaking at less than ~\$130–\$135.

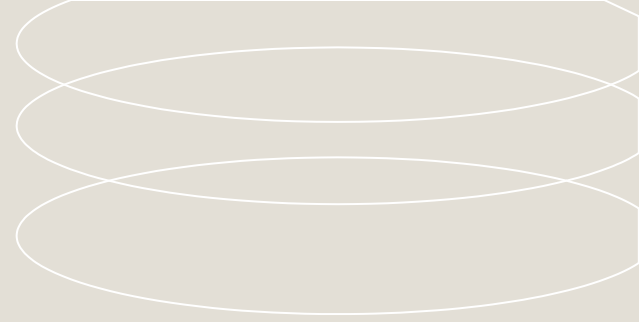
Contrary to prior literature, the performance of lower similarity filings (Bins 1–3) was mixed; Bin 3 underperformed, while Bin 1 showed moderate performance, and Bins 2 and 4 performed strongly.

Insights

While our bin-based portfolios showed distinct return patterns over time, the correlation matrix reveals no strong linear relationship between cosine similarity and short-term returns. This divergence suggests that similarity's impact on performance may be conditional on other firm characteristics — echoing the nuanced findings of the original paper.



Conclusions



- Firms with high 10-K similarity (Bin 5) consistently outperformed, confirming findings from Cohen et al. (2020).
- Bin 5 portfolios grew the most over 20 years, reaching \$150+, suggesting market preference for consistent, standardized disclosures.
- Low similarity bins (e.g., Bin 3) showed weaker or inconsistent returns, indicating that novelty in filings may signal risk or uncertainty.
- Despite some anomalies (e.g., strong performance in Bin 2), the overall trend supports a positive link between similarity and returns.
- Our analysis highlights the predictive power of textual similarity in financial disclosures for equity performance.



Questions?

Or website suggestions?

