Ask EDGAR:

Insights into Asset Management from Big Data and Natural Language Processing

Simona Abis and Anton Lines

Columbia University

NBER Summer Institute

Big Data and High-Performance Computing for Financial Economics

Capital Allocation

- Understanding capital allocation is key to understand the functioning of market economies

- Mutual funds are often used as a laboratory to understand capital allocation
  - Prior research delivered low explanatory power when explaining capital flows using only "hard information"
    (e.g. TNA, fees, past return)
  - Investors also have access to "soft information"
    (e.g. marketing materials, regulatory disclosures, in-person meetings)
  - Prospectuses are the main document containing funds information; likely proxy for other forms of communication
  - The SEC has been urging investors to be weary of past performance and to read prospectuses carefully when making investment decisions
    (Rule 156)
  - Disclosure requirements have been shown to be very costly for funds

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information
- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents
- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information

- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents

- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information

- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents

- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information

- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents

- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information

- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents

- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Questions

- QUESTION 1:
  Can we extract informative signals from the text of mutual fund prospectuses above and beyond what can be learned by analyzing hard information?
  - Advances in computing power, machine learning and NLP allow for collection, storage and detailed analysis of textual information

- QUESTION 2:
  Are investors following SEC guidelines and paying attention to this information in their capital allocation decisions?
  - Investors are limited in their capacity for processing information while being overwhelmed with a wealth of legal documents

- QUESTION 3:
  Are they doing so in an efficient manner?
  - Investors vary in sophistication hence they might be looking for different information/interpreting the same information differently

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US actively managed mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be used as quasi-natural experiments

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US active equity mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be used as quasi-experiments

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US active equity mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be used as quasi-natural experiments

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US equity mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be used as quasi-natural experiments

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US active equity mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be exploited for identification

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - Linguistic and clustering analysis of US active equity mutual funds
    - Defined empirical strategy
    - Identified regulatory changes to be exploited for identification

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - ⋆ Linguistic and clustering analysis of US active equity mutual funds
    - ⋆ Defined empirical strategy
    - ⋆ Identified regulatory changes to be exploited for identification

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - ⋆ Linguistic and clustering analysis of US active equity mutual funds
    - ⋆ Defined empirical strategy
    - ⋆ Identified regulatory changes to be exploited for identification

Research Plan

- DATA:
  Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
  - Status: Finalizing comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of all historical filings

- THEORY:
  Develop a theoretical framework based on rational inattention to guide data analysis
  - Status: Work in progress (model set-up and intuition)

- EMPIRICS:
  Test hypotheses empirically through machine learning analysis of funds' prospectuses, using regulatory changes as quasi-natural experiments
  - Status:
    - ⋆ Linguistic and clustering analysis of US active equity mutual funds
    - ⋆ Defined empirical strategy
    - ⋆ Identified regulatory changes to be exploited for identification

Literature

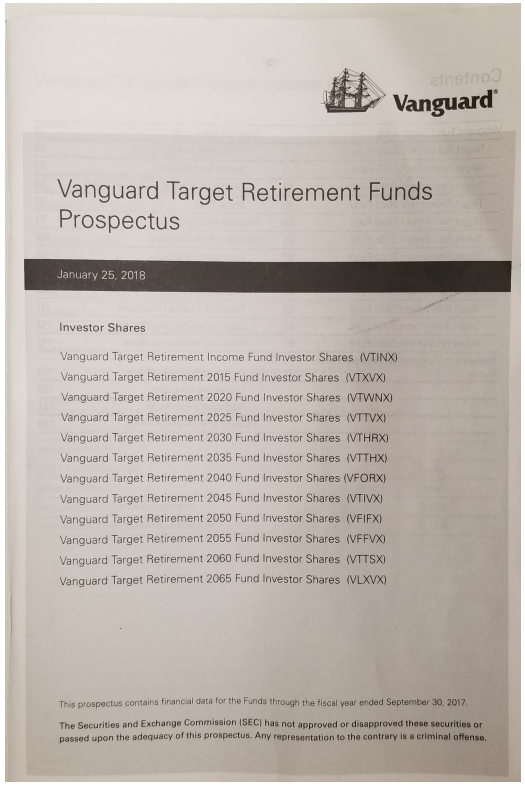
- Determinants of mutual fund flows: Sirri & Tufano (1998), Jain & Wu (2000), Del Guercio & Tkac (2002), Barber, Odean & Zheng (2003), Berk & Green (2004), Del Guercio & Tkac (2008), Ivković & Weisbenner (2009), Gennaioli Shleifer & Vishny (2015)
  - We'll provide new insights using variables extracted from soft information
- Rational inattention: Sims (2003); Mackowiak and Wiederholt (2009, 2015); Van Nieuwerburgh and Veldkamp (2010); Kacperczyk, Van Nieuwerburgh and Veldkamp (2016)
  - We'll use this learning framework to explain capital allocation decisions of investors with varying levels of sophistication
- Textual analysis and machine learning: Subramanian, Insley & Blackwell (1993), Philpot & Johnson (2007), Tetlock (2007), Tetlock, Saar-Tsechansky, Macskassy (2008), Manela and Moreira (2017), Abis (2018), Ryans (2018), Kelley, Manela and Moreira (2018)
  - We'll apply supervised and unsupervised learning to extract signals from mutual fund prospectuses
- SEC regulatory changes and EDGAR usage: Johnson (2004), Agarwal, Mullally, Tang & Yang (2015), Agarwal, Vashishtha & Venkatachalam (2017), Gao and Huang (2018)
  - We'll exploit regulatory changes to study the reaction of flows to cross-sectional differences in mandatory disclosures

Roadmap

Roadmap

Prospectuses Availability

- Funds are required to publish prospectuses regularly
- There are clear guidelines regarding the information these should contain
  - Funds can be sued by the SEC for misrepresenting their behavior
- Prospectuses are publicly available through the EDGAR system since 1995
  - Sophisticated investors can automate access using the FTP of the SEC
  - Retail investors might also access prospectuses of selected funds through their online brokerage accounts or by post



Prospectuses Description

- They are divided in sections addressing different regulatory questions e.g.:
  - Principal Investment Strategies (PIS)
  - Principal Risks (PIR)
- The content, writing style and length of different sections vary substantially
  - Crucial to condition on sections when comparing text cross-sectionally
- Regulatory requirements can be satisfied in just a few sentences
  - Some funds choose to write substantially more

E.g.: Vanguard - JAG Large Cap Growth Fund

Principal Investment Strategies

The Fund invests primarily in common stocks of U.S. companies that the Fund's advisor believes have strong earnings and revenue growth potential. Under normal conditions, the Fund will invest at least 80% of the Fund's net assets plus any borrowings for investment purposes in large cap stocks defined as stocks of companies with market capitalizations of at least $8 billion.

The advisor's employs a bottom-up, quantitatively-derived buy discipline to identify stocks the advisor believes have superior earnings and revenue growth characteristics. The cornerstone of the advisor's investment process is a proprietary multi-factor model that scores several thousand equity securities according to a variety of weighted factors measuring earnings and revenue growth, valuation, size and relative strength. The sell discipline is designed to eliminate portfolio holdings with inferior price performance and deteriorating earnings and revenue growth factors.

The Fund actively trades its portfolio investments, which may lead to higher transaction costs that may affect the Fund's performance.

Principal Risks of Investing in the Fund

As with any mutual fund, there is no guarantee that the Fund will achieve its objective. Investment markets are unpredictable and there will be certain market conditions where the Fund will not meet its investment objective and will lose money. The Fund's net asset value and returns will vary and you could lose money on your investment in the Fund and those losses could be significant.

The following summarizes the principal risks of investing in the Fund. These risks could adversely affect the net asset value, total return and the value of the Fund and your investment.

- **Equity Securities Risks.** Common stocks are subject to market risks that affect the value of the Fund. Factors such as interest rate levels, market conditions, and political events may adversely affect equity prices.

- **Management Risk.** The Portfolio Manager's judgments about the attractiveness, value and potential appreciation of particular stocks, options or other securities in which the Fund invests or sells short may prove to be incorrect and there is no guarantee that the Portfolio Manager's

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - "Principal Investment Strategies" OR "Principal Strategies" OR "Principal Investments and Strategies of the Fund" OR ...
  - Files don't all follow the same standard
    - .pdf, different .html standards, .txt, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - "Principal Investment Strategies" OR "Principal Strategies" OR "Principal Investments and Strategies of the Fund" OR ...
  - Files don't all follow the same standard
    - .pdf, different .html standards, .txt, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - "Principal Investment Strategies" OR "Principal Investment Strategy" OR "Principal Investments and Strategies" OR "Principal Investment Strategies of the Fund" OR ...
  - Files don't all follow the same standard
    - .pdf, different .html standards, .txt, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - "Principal Investment Strategies" OR "Principal Investment Strategy" OR
    - "Principal Investments and Strategies" OR "Principal Strategies"
  - Files don't all follow the same standard
    - .pdf, different .html standards, .txt, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - "Principal Investment Strategies" OR "Principal Investment Strategy" OR
    - "Principal Investments and Strategies" OR "Principal Strategy"
  - Files don't all follow the same standard
    - .pdf, different .html standards, .txt, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Challenges

- The EDGAR Mutual Fund database includes over 1 million prospectuses for a total size of $\approx 19 TB$
- Our unit of analysis is the individual fund but prospectuses can only be searched by the name or CIK code of their fund family
  - Doesn't easily map to known identifiers
- Fund families might publish the prospectus of multiple funds in the same file
  - Challenging to automatically identify parts belonging to funds of interest
- Files don't always follow the same structure, particularly historically:
  - Section names might vary by fund e.g.
    - Principal Investment Strategies OR Principal Strategies OR Principal Investments and Strategies of the Fund OR ...
  - Files don't all follow the same standard
    - pdf, different html standards, xbrl, ...
  - Most files are organized by fund but some follow a "functional" structure

Our Parser

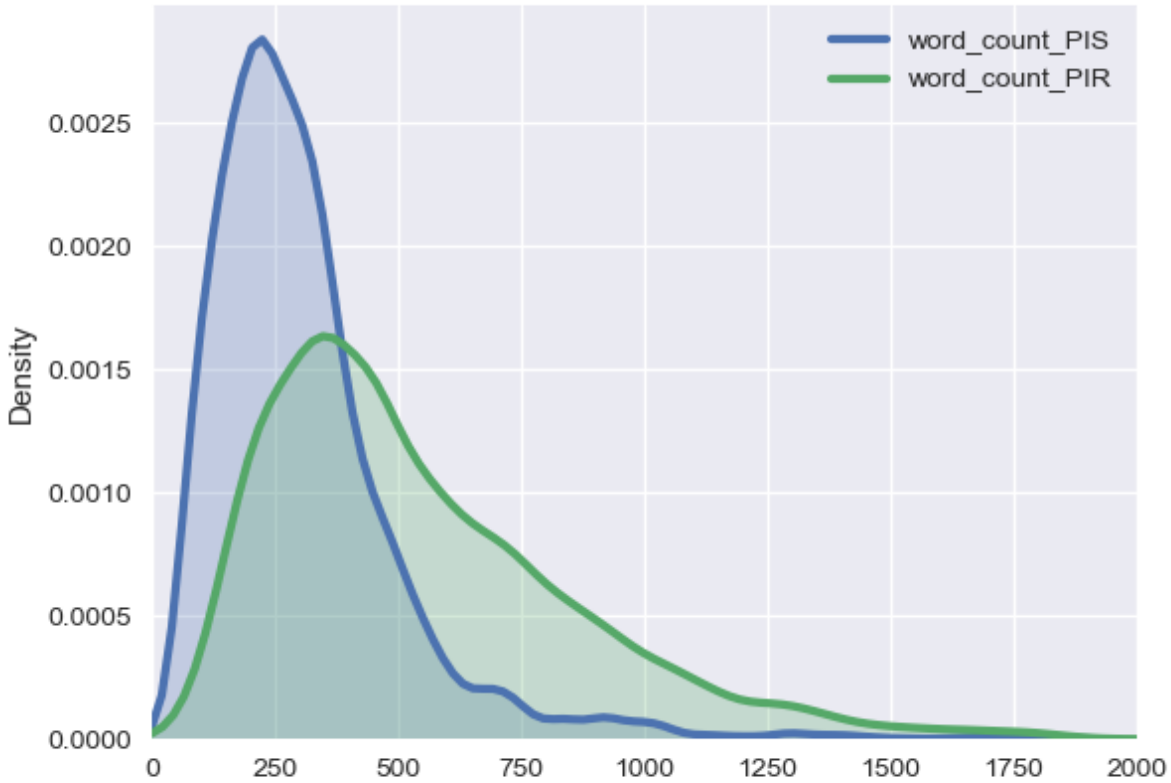
- The Ask EDGAR project is fully coded in Python
- The parsing job has so far been applied to US active equity mutual funds
- Steps:
  - Download file indices using the SEC FTP
  - Download filings of interest using URLs obtained from the indices
  - Create a mySQL database to store all raw and parsed data
  - Apply a comprehensive parser to all downloaded filings as follows:
    - Identify list of funds present in files & verify overlap with funds of interest
    - Look for xbrl-style tags, if available parse entire file into sections
    - Otherwise use "Beautiful Soup" to identify sections of interest
    - Finally use "PyParsing" & "Regex" when there isn't a clear html structure
  - The output table includes:
    - Fund name, identifier and month of filing
    - A separate variable for each section of the prospectus

Roadmap

PIS vs. PIR

- Strategy descriptions (PIS) are substantially shorter than Risk ones (PIR)
- But they are harder to understand - Dale Chall Score:

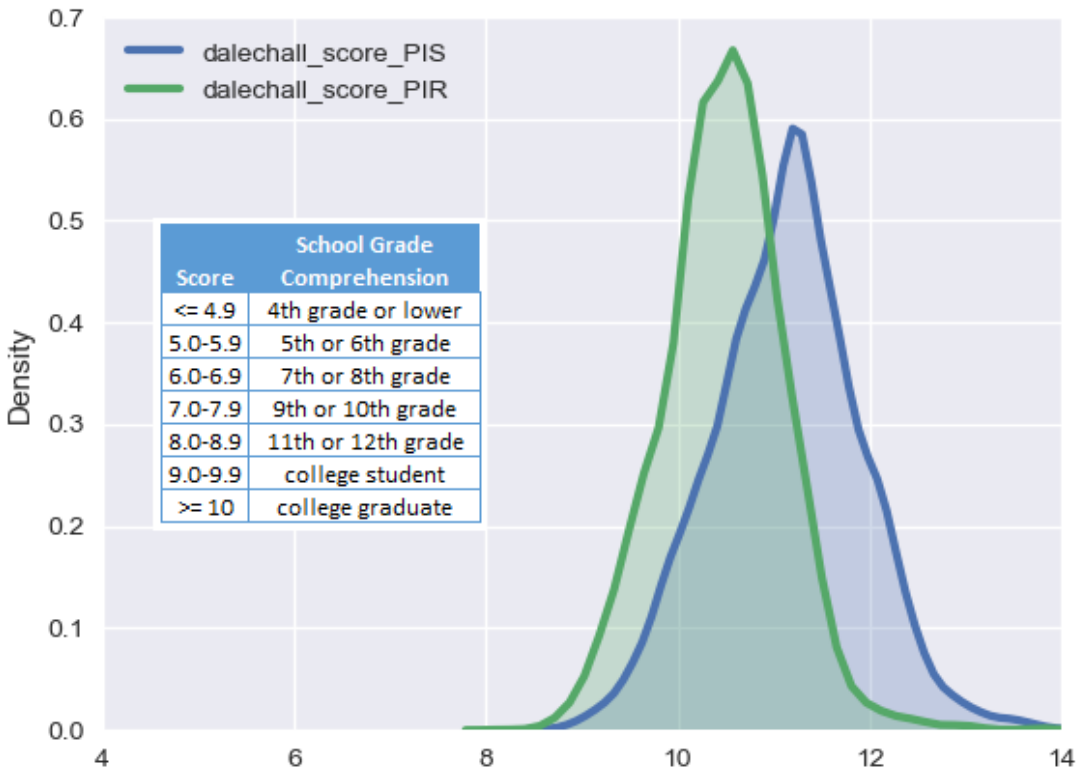
$$\begin{cases} 3.6365 & \frac{nonDaleChalCount}{wordCount} > 0.5 \\ 0 & otherwise \end{cases} + 15.79 * \frac{nonDaleChalCount}{wordCount} + 0.0496 \frac{wordCount}{sentCount}$$



PIS vs. PIR

- Strategy descriptions (PIS) are substantially shorter than Risk ones (PIR)
- But they are harder to understand - Dale Chall Score:

$$\begin{cases} 3.6365 & \frac{nonDaleChalCount}{wordCount} > 0.5 \\ 0 & otherwise \end{cases} + 15.79 * \frac{nonDaleChalCount}{wordCount} + 0.0496 \frac{wordCount}{sentCount}$$

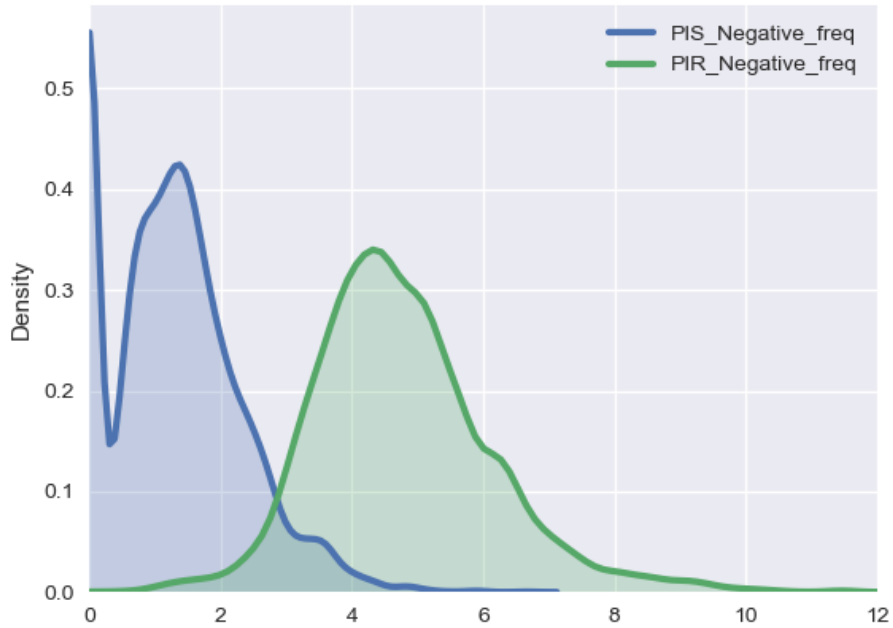


Score	School Grade Comprehension
<= 4.9	4th grade or lower
5.0-5.9	5th or 6th grade
6.0-6.9	7th or 8th grade
7.0-7.9	9th or 10th grade
8.0-8.9	11th or 12th grade
9.0-9.9	college student
>= 10	college graduate

PIS vs. PIR (continued)

Using Loughran and McDonald sentiment word lists we find that PIR contain:

- A higher frequency of Negative words
- A higher frequency of Uncertainty and Litigious and Constraining words
  A lower frequency of Positive words



PIS vs. PIR (continued)

Using Loughran and McDonald sentiment word lists we find that PIR contain:

- A higher frequency of Negative words
- A higher frequency of Uncertainty and Litigious and Constraining words
  A lower frequency of Positive words

$$\text{Sentiment}_{i,t} = \delta T_{i,t} + v_i + u_t + \epsilon_{i,t}$$

where: i=fund; t=time

$$\text{T}_{i,t} = \begin{cases} 0 & \text{if } SectionType_{i,t} = PIS \\ 1 & \text{if } SectionType_{i,t} = PIR \end{cases}$$

	(1) negative	(2) positive	(3) litigious	(4) uncertainty	(5) constraining
T	3.495***	-0.505***	0.495***	0.953***	0.252***
	(58.34)	(-12.29)	(20.83)	(19.79)	(16.86)
N	24716	24716	24716	24716	24716

t statistics in parentheses

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

Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

Clustering

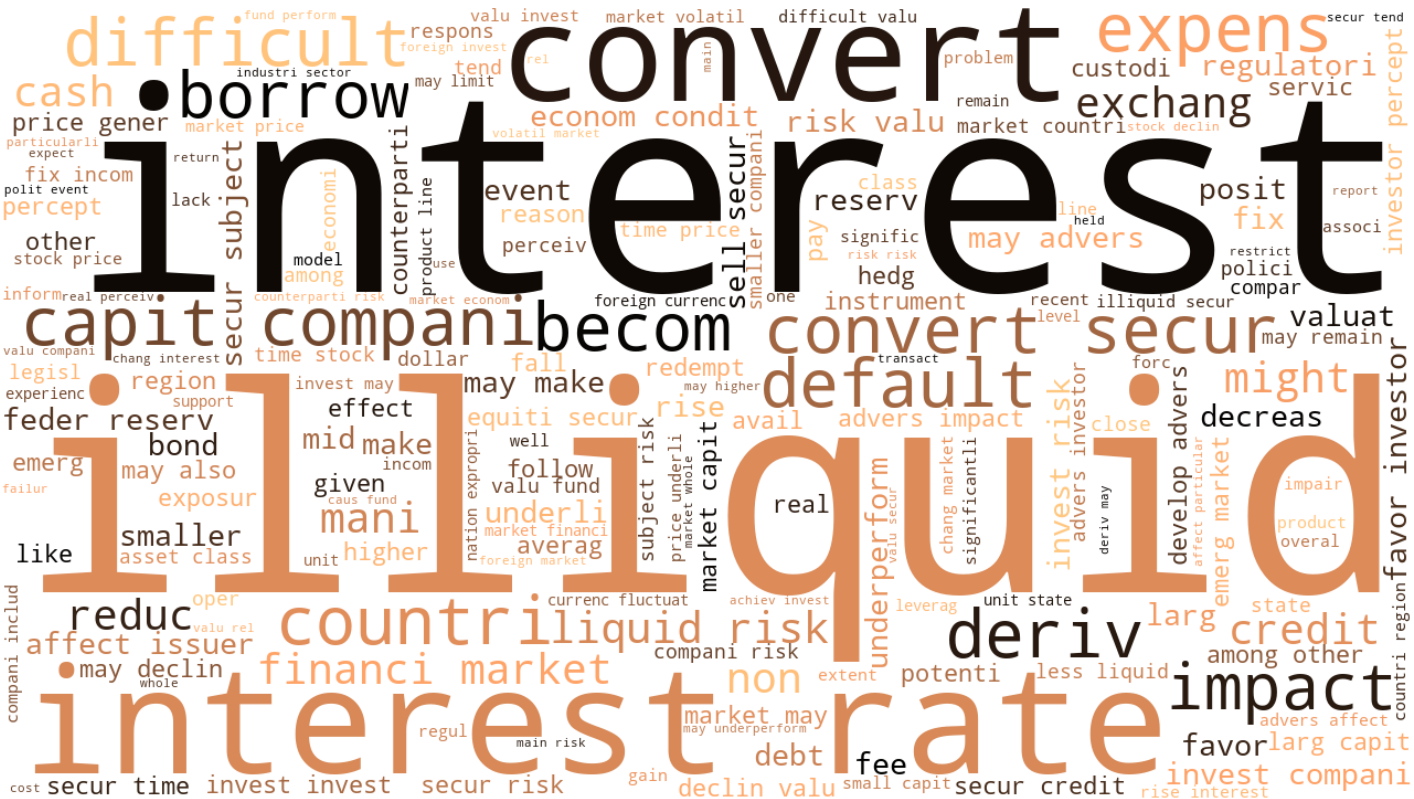
- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIR Derivatives

Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIR Illiquidity



Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIR Technology

Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIR Regulatory

Clustering

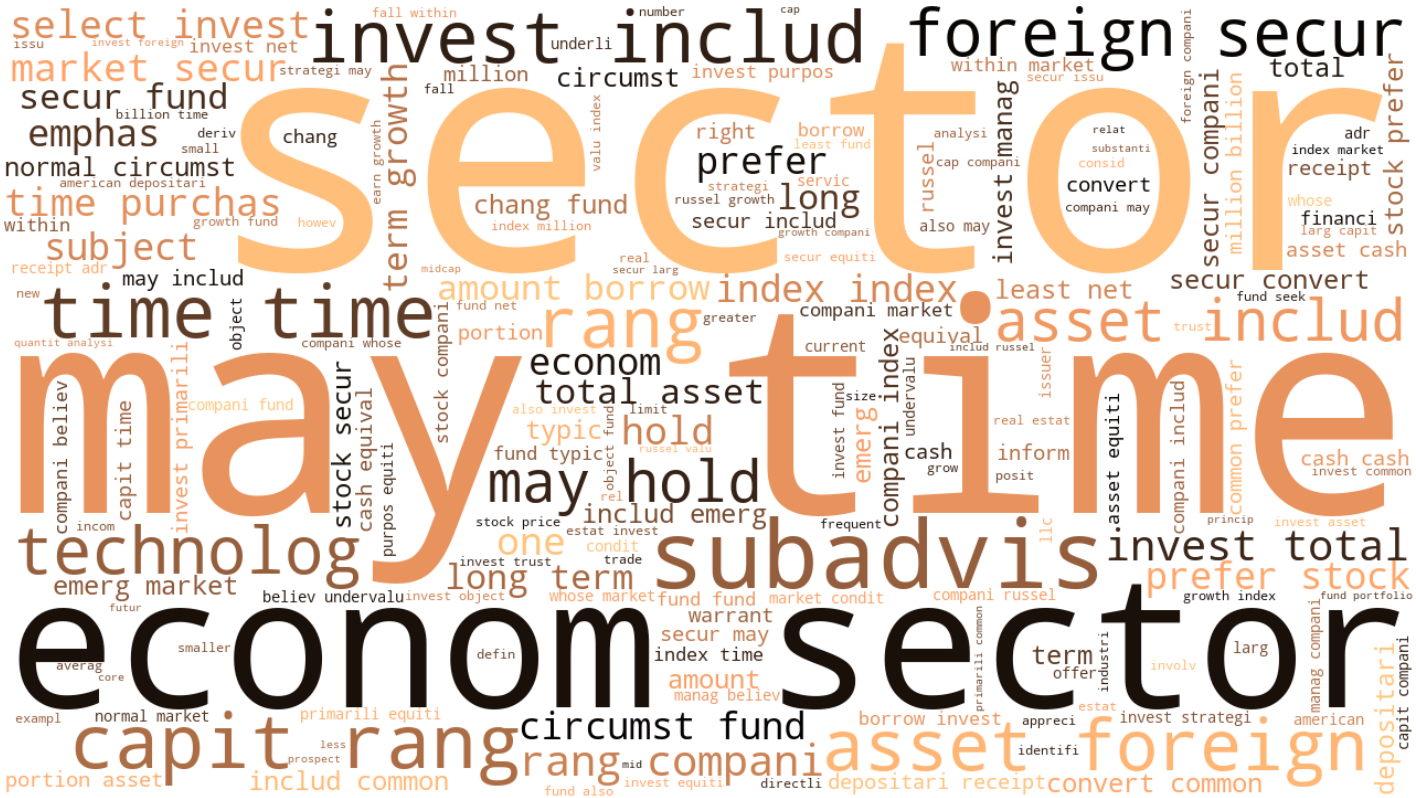
- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIR Foreign

Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

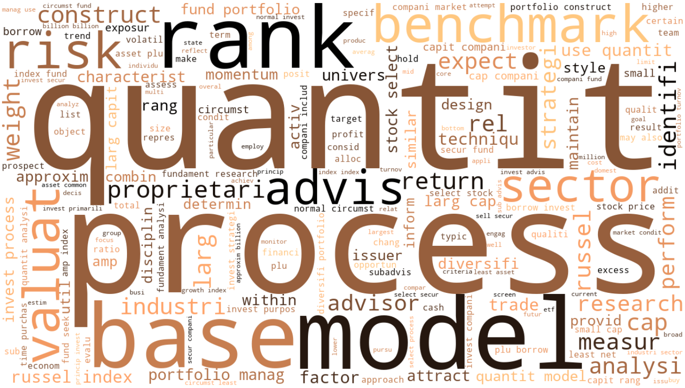
PIS Timing



Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIS Quant



Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIS Macro

Clustering

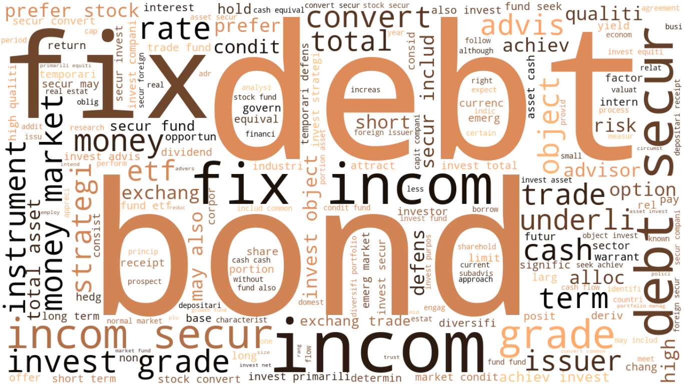
- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIS Long Term

Clustering

- We used unsupervised learning to group descriptions into clusters
  - Gaussian Mixture: Clusters are characterized by Gaussians over the vector space of word and bi-gram relative frequencies
  - PIS was best described by 13 clusters, PIR by 7 e.g:

PIS Fixed Income



Roadmap

Set-up

- Timeline:
  - $t = 1$: investors allocate learning capacity
  - $t = 2$: investors choose portfolio allocation
  - $t = 3$: assets pay off

- Assets: one riskless asset, one risky benchmark and $n - 1$ mutual funds
  - Benchmark: $f_b = b$ with $b \sim N(\mu_b, \sigma_b^2)$
  - Funds $i = 1, \ldots, n$ with payoff: $f_i = \alpha_i - D_i + b$
    - $\alpha_i = z_i - g(Q_i)$; $Q_i = \int_0^1 q_{ij} dj$;
    - $z_i \sim N(\mu_{z,i}, \sigma_{z,i}^2)$; $\sum_i \mu_{z,i} = 0$
    - $D_i = \begin{cases} 0 & \text{with probability } (1 - d_i) \\ D & \text{with probability } d_i \end{cases}$
    - $D \sim N(\mu_d, \sigma_d^2)$; $d_i \ll 1$ for each i
  - Assets are in perfectly elastic supply and borrowing is unrestricted
  - Prices and fund abilities are exogenously given

Set-up (continued)

- Learning: investors observe signals about:
  - Funds ability: $s_{ij}^z = z_i + \epsilon_{ij}^z$ *with* $\epsilon_{ij}^z \sim N(0, \sigma_{z,ij}^2)$
  - Loss due to untrustworthy behavior: $s_{ij}^d = D_i + \epsilon_{ij}^d$ *with* $\epsilon_{ij}^d \sim N(0, \sigma_{z,ij}^2)$
  - Learning determines $\sigma_{z,ij}^2$ and $\sigma_{d,ij}^2$

- Investors: unit-mass of mean-variance investors with risk aversion ρ, initial wealth W_0 and learning capacity K
  - Financially literate investors (mass χ):
    - ⋆ Learn about funds ability z_i and the common loss D s.t. $\sum_i \left(\sigma_{z,ij}^2\right)^{-1} + \left(\sigma_d^2\right)^{-1} = K$
    - ⋆ Know the probabilities of untrustworthy behavior of all funds d_i
  - Financially illiterate investors (mass $1 - \chi$):
    - ⋆ Do not receive a signal about funds ability z_i
    - ⋆ Learn about loss due to untrustworthy behavior D_i s.t. $\sum_i \left(\sigma_{d,ij}^2\right)^{-1} = K$

Intuition

- Model solution: Investors' optimal allocations of learning capacity and capital across funds as a function of signals precision and aggregate allocations:

$$\{K_{ij}^*(\hat{\Sigma}_{z,j}, \hat{\Sigma}_{d,j}, g(Q_i^*)),\ q_{ij}^*(\hat{\Sigma}_{z,j}, \hat{\Sigma}_{d,j}, g(Q_i^*))\}$$

  - Investors' capital allocation also influence the distribution of α through diminishing returns to scale of strategies: $\alpha_i = z_i - g(Q_i)$; $Q_i = \int_0^1 q_{ij} dj$
- Why this set-up?
  - Model capital allocation with heterogeneity in learning by investors
    - ⋆ Different sections of prospectuses can appeal to investors with varying levels of financial sophistication
      - e.g: PIR easier to understand $\rightarrow$ learning about D_i?
    - ⋆ Learning substitution effects can lead to counter-intuitive allocations
      - e.g: Sophisticated investors look for funds with high ability & tail risk?
    - ⋆ Changes in availability/clarity of information can impact capital allocation
      - e.g: Does better information always increase efficiency?

Empirical Strategy

- Measurement
  - Identify textual proxies for trustworthiness/tail-risk signals by training predictive algorithms on SEC lawsuits and ex-post risk shifting behavior
    - ⋆ Supervised learning e.g. SVM, Random Forest
    - ⋆ Examine their ability to predict flows of unsophisticated investors
  - Identify textual features associated with performance
    - ⋆ Supervised, unsupervised and/or reinforcement learning methods
    - ⋆ Examine their ability to predict flows of sophisticated investors
- Identification
  - Relate textual measures to fund growth one year after inception
  - Regulatory changes:
    - ⋆ 1995: introduction of online EDGAR distribution system
    - ⋆ 1998 (Rule 421): Readability act
    - ⋆ 1999: Increase disclosure requirements in PIS
    - ⋆ 2004: More frequent disclosure

Roadmap

Conclusion

- Very preliminary ambitious work!
  - Any feedback/criticism is very welcome
- Wealth of information in EDGAR filings largely unexploited, particularly in relation to asset management
  - Machine learning methods are necessary to fully exploit its potential
- Learning perspective on asset allocation
  - Complements existing literature
    - ⋆ Berk & Green (2004), Gennaioli Shleifer & Vishny (2015)