ASSET EMBEDDINGS

Xavier Gabaix   Ralph Koijen   Robert Richmond   Motohiro Yogo

Harvard - Chicago - NYU - Princeton

August 2024

IDENTIFYING SIMILAR FIRMS

- ▶ In economics, we often try to find similar firms or assets.
  - ▶ E.g., similar growth rates, expected returns, risk, asset substitution, product markets, ...

- ▶ Common practice: Use observable characteristics.
  - ▶ E.g., industry definitions, accounting data, ...

IDENTIFYING SIMILAR FIRMS

- ▶ In economics, we often try to find similar firms or assets.
  - ▶ E.g., similar growth rates, expected returns, risk, asset substitution, product markets, ...

- ▶ Common practice: Use observable characteristics.
  - ▶ E.g., industry definitions, accounting data, ...

- ▶ Those characteristics may be quite imperfect.
  - ▶ Standardized accounting data are an incomplete summary.
    - ▶ E.g., number of subscribers at Netflix, ...
  - ▶ New economic environments call for creative, new characteristics.
    - ▶ E.g., exposure to COVID-19, intangibles or AI.

IDENTIFYING SIMILAR FIRMS

- In economics, we often try to find similar firms or assets.
  - E.g., similar growth rates, expected returns, risk, asset substitution, product markets, ...

- Common practice: Use observable characteristics.
  - E.g., industry definitions, accounting data, ...

- Those characteristics may be quite imperfect.
  - Standardized accounting data are an incomplete summary.
    - E.g., number of subscribers at Netflix, ...
  - New economic environments call for creative, new characteristics.
    - E.g., exposure to COVID-19, intangibles or AI.

- This paper: Use asset embeddings to measure firm similarity.

WHAT ARE EMBEDDINGS?

▶ Embeddings: Represent data (e.g., words) as vectors in a potentially high-dimensional space: $x_a \in \mathbb{R}^K$.

▶ Embeddings play a central role in the development of large language models (LLMs).

▶ In LLMs, embeddings capture the similarity between words and it allows us to do "math with words:

$$x_{\mathsf{Paris}} - x_{\mathsf{France}} + x_{\mathsf{Spain}} \simeq x_{\mathsf{Madrid}}.$$

WHAT ARE EMBEDDINGS?

▶ Embeddings: Represent data (e.g., words) as vectors in a potentially high-dimensional space: $x_a \in \mathbb{R}^K$.

▶ Embeddings play a central role in the development of large language models (LLMs).

▶ In LLMs, embeddings capture the similarity between words and it allows us to do "math with words:

$$x_{\text{Paris}} - x_{\text{France}} + x_{\text{Spain}} \simeq x_{\text{Madrid}}.$$

▶ Embedding vectors are learned from (lots of) data (not preselected).

▶ Despite the success of embedding techniques in these fields, their application in finance and economics largely unexplored.

IDEAL DATA TO ESTIMATE EMBEDDINGS?

- ► We introduce the concept of asset embeddings.
  - ► A vector representation for each asset, that we learn from data.

- ► Which data to use?

IDEAL DATA TO ESTIMATE EMBEDDINGS?

▶ We introduce the concept of asset embeddings.
  ▶ A vector representation for each asset, that we learn from data.

▶ Which data to use?

▶ Our answer: Just like
  ▶ documents organize words in language modeling,
  ▶ images organize pixels in computer vision,
  ▶ songs organize notes in audio,

  investors organize assets in finance and economics.

IDEAL DATA TO ESTIMATE EMBEDDINGS?

▶ We introduce the concept of asset embeddings.
  ▶ A vector representation for each asset, that we learn from data.

▶ Which data to use?

▶ Our answer: Just like
  ▶ documents organize words in language modeling,
  ▶ images organize pixels in computer vision,
  ▶ songs organize notes in audio,

  investors organize assets in finance and economics.

▶ Theoretically, we show how embeddings can be recovered by "inverting the asset demand system."

WHICH METHOD TO LEARN EMBEDDINGS?

► Which method to use?

WHICH METHOD TO LEARN EMBEDDINGS?

▶ Which method to use?

▶ Traditional approach: LSA (Latent Semantic Analysis), which is related to PCA/recommender systems.

▶ The recent ML/AI literature went way beyond that.

  ▶ Context-invariant embeddings: E.g., GloVe and Word2Vec.

  ▶ Embeddings with context: E.g., transformer models (e.g., BERT and GPT).

  ▶ Parameters are estimated using masked language modeling.

INVESTOR EMBEDDINGS

▶ Even though our focus is on asset embeddings, we obtain investor embeddings as a by-product: $\lambda_i \in \mathbb{R}^K$.

  ▶ Learned vector representations of each investor's "taste for characteristics"

INVESTOR EMBEDDINGS

- ▶ Even though our focus is on asset embeddings, we obtain investor embeddings as a by-product: $\lambda_i \in \mathbb{R}^K$.
    - ▶ Learned vector representations of each investor's "taste for characteristics"

- ▶ Examples of applications:
    - ▶ Classify investors beyond institutional type, size, and activeness.
    - ▶ Identify crowded trades.
    - ▶ Performance measurement (extending Daniel, Grinblatt, Titman, and Wermers, 1997).

FIVE MAIN CONTRIBUTIONS

1. Micro-found the use of holdings data as embeddings data (in the paper).

2. Three benchmarks to compare asset embedding models.
   - Building on the success of benchmark in AI (e.g., ImageNet).

3. Explore different modeling architectures to learn asset embeddings based on language models.

4. Evaluate benchmarks for asset embeddings, text-based embeddings, and observed characteristics.

5. Use earnings calls data to interpret the embeddings.
   - Extends to any other form of text data (e.g., WSJ articles, analyst reports, ...).

RECOMMENDER SYSTEMS

▶ Recommender systems, with $\theta = (x_a, \lambda_i, \delta_i, \delta_a)$,

$$\min_\theta \frac{1}{IA} \sum_{i,a} (h_{ia} - \delta_i - \delta_a - \lambda_i' x_a)^2 + \frac{\xi}{IK} \sum_i \lambda_i' \lambda_i + \frac{\xi}{AK} \sum_a x_a' x_a,$$

where
  ▶ h_{ia}: Log holdings.
  ▶ x_a: Asset embeddings.
  ▶ λ_{iq}: Investor embeddings.

▶ Analogous to LSA in the NLP literature.[1]

---

[1]Dumais, Furnas, Landauer, and Deerwester (1988).

IMPLEMENTATIONS OF RECOMMENDER SYSTEMS

▶ To understand how to best extract information from holdings, we consider five variants:

1. Binary, $\mathbb{I}_{H_{ia}>0}$.

2. Percentile ranks of H_{ia} with missing values set to zero.

3. h_{ia} with missing values set to zero.

4. h_{ia} with missing values set to the smallest active position.

5. h_{ia} using only the non-missing values.

WORD2VEC

- General approach to estimate language models, such as Word2Vec,[2]
  - Task: Guess masked words.
    - E.g. "Please pass me the _____ and pepper".
  - Use a context window to maximize the probability of a missing word given the context info:

$$\mathbb{P}(w_a \mid w_c) = \frac{\exp(x_a' x_c)}{\sum_b \exp(x_b' x_c)}.$$

[2]Mikolov, Sutskever, Chen, Corrado, Dean (2013a, b).

WORD2VEC

- General approach to estimate language models, such as Word2Vec,[2]
  - Task: Guess masked words.
    - E.g. "Please pass me the _____ and pepper".
  - Use a context window to maximize the probability of a missing word given the context info:

$$\mathbb{P}(w_a \mid w_c) = \frac{\exp(x'_a x_c)}{\sum_b \exp(x'_b x_c)}.$$

- Estimation using holdings data:
  - Sentences $\Rightarrow$ Investors.
  - Words $\Rightarrow$ Assets.
  - Objective: Guess masked assets (cross entropy).

---

[2]Mikolov, Sutskever, Chen, Corrado, Dean (2013a, b).

MASKED ASSET MODELING

▶ Example: The ARKK ETF in July 2023:

Holdings Data – ARKK
As of 07/07/2023

ARK INVEST

ARKK
ARK Innovation ETF

	Company	Ticker	CUSIP	Shares	Market Value ($)	Weight (%)
1	TESLA INC	**TSLA**	88160R101	3,496,872	$967,024,982.88	12.43%
2	COINBASE GLOBAL INC -CLASS A	**COIN**	19260Q107	7,945,138	$620,515,277.80	7.98%
3	ROKU INC	**ROKU**	77543R102	8,865,426	$546,110,241.60	7.02%
4	~~ZOOM VIDEO COMMUNICATIONS-A~~	~~ZM~~	~~98980L101~~	~~8,258,591~~	~~$534,248,251.79~~	~~6.87%~~
5	UIPATH INC - CLASS A	**PATH**	90364P105	28,152,366	$463,106,420.70	5.95%
6	BLOCK INC	**SQ**	852234103	7,069,493	$456,759,942.73	5.87%
7	EXACT SCIENCES CORP	**EXAS**	30063P105	4,031,264	$368,739,718.08	4.74%
8	UNITY SOFTWARE INC	**U**	91332U101	8,350,868	$338,627,697.40	4.35%
9	SHOPIFY INC - CLASS A	**SHOP**	82509L107	5,430,238	$335,751,615.54	4.32%
10	DRAFTKINGS INC-CL A	**DKNG UW**	26142V105	12,035,607	$303,658,364.61	3.90%

CONTEXT AND SELF-ATTENTION: A SIMPLE EXAMPLE

- ► So far, we have one x_a per asset, say, Apple, with no context.
- ► How does attention[3] work?

---

[3]Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin (2017).

CONTEXT AND SELF-ATTENTION: A SIMPLE EXAMPLE

▶ So far, we have one x_a per asset, say, Apple, with no context.

▶ How does attention[3] work?

1. $\mathcal{H}_i$: Stocks in the portfolio of manager i.

2. For stock $a \in \mathcal{H}_i$, compute a similarity score with the other stocks $b \in \mathcal{H}_i$

$$\sigma_{ab} = x_a' x_b.$$

   x_a: Query.
   x_b: Key.

3. Compute the contextualized embedding, x_a^i,

$$x_a^i = \sum_{b \in \mathcal{N}_i} \frac{e^{\sigma_{ab}}}{\sum_{c \in \mathcal{N}_i} e^{\sigma_{ac}}} x_b.$$

   x_b: Value.

[3]Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin (2017).

GENERALIZING ATTENTION: TRANSFORMERS

▶ Transformer models generalize this idea.
  ▶ Query: $q_a = W^Q x_a$.
  ▶ Key: $k_a = W^K x_a$.
  ▶ Value: $v_a = W^V x_a$.

▶ The contextualized embedding is then computed as

$$x_a^i = \sum_{b \in \mathcal{N}_i} \frac{e^{\sigma_{ab}}}{\sum_{c \in \mathcal{N}_i} e^{\sigma_{ac}}} v_b, \qquad \sigma_{ab} = q_a' k_b.$$

▶ The matrices W_Q, W_K, and W_V are learned from (lots of) data and determine which aspects of the context are important.

GENERALIZING ATTENTION: TRANSFORMERS

- ▶ Transformer models generalize this idea.
    - ▶ Query: $q_a = W^Q x_a$.
    - ▶ Key: $k_a = W^K x_a$.
    - ▶ Value: $v_a = W^V x_a$.
- ▶ The contextualized embedding is then computed as

$$x_a^i = \sum_{b \in \mathcal{N}_i} \frac{e^{\sigma_{ab}}}{\sum_{c \in \mathcal{N}_i} e^{\sigma_{ac}}} v_b, \qquad \sigma_{ab} = q_a' k_b.$$

- ▶ The matrices W_Q, W_K, and W_V are learned from (lots of) data and determine which aspects of the context are important.

- ▶ Features of the full AssetBERT model
    - ▶ Stack multiple attention layers with multi-headed attention.
    - ▶ Add a feed-forward layer between each self-attention layer:

    $$FF(x) = \max(0, xW_1 + b_1)W_2 + b_2,$$

    where the dimensionality of the inner layer $\gg \dim(x)$.
    - ▶ Add position embeddings.

DATA

- ▶ Holdings data from FactSet:
  - ▶ Hedge funds, mutual funds, ETFs, closed-end funds, variable annuity funds.

- ▶ Sample construction:
  - ▶ 2005.Q1 - 2022.Q4.
  - ▶ Remove nano and micro caps.
  - ▶ Keep investors (stocks) with at least 20 positions (investors).

- ▶ Accounting data and stock returns from CRSP / Compustat, using the Jensen, Kelly, and Pedersen (2023) construction.

- ▶ Earnings calls data from FactSet.
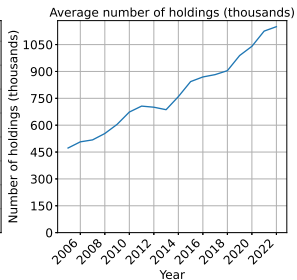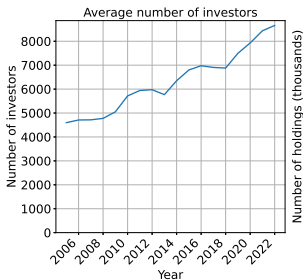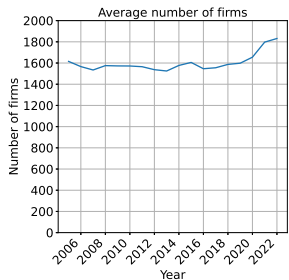
▶ Main takeaway:
  ▶ The number of holdings per firm steadily increased over time.

THREE BENCHMARKS

- ▶ Benchmark competitions identify the best performing models in AI and give metrics for success.
    - ▶ E.g., ImageNet in computer vision.

1. Predicting relative valuations.
    - ▶ Decompose $m_a = \beta_0 + \beta_1 b_{at} + m_a^{\perp}$.
    - ▶ Estimate $m_a^{\perp} = \gamma_0 + \gamma_1' x_a + \epsilon_a$ on 80% of the sample.
    - ▶ Evaluate using the R^2 on the remaining 20%.

THREE BENCHMARKS

- ▶ Benchmark competitions identify the best performing models in AI and give metrics for success.
    - ▶ E.g., ImageNet in computer vision.

1. Predicting relative valuations.
    - ▶ Decompose $m_a = \beta_0 + \beta_1 b_{at} + m_a^\perp$.
    - ▶ Estimate $m_a^\perp = \gamma_0 + \gamma_1' x_a + \epsilon_a$ on 80% of the sample.
    - ▶ Evaluate using the R^2 on the remaining 20%.

2. Predicting missing characteristics.
    - ▶ Estimate $c_a = \delta_0 + \delta_1' x_a + \epsilon_a$ on 80% of the sample.
    - ▶ Evaluate using the R^2 on the remaining 20%.

THREE BENCHMARKS

- ▶ Benchmark competitions identify the best performing models in AI and give metrics for success.
  - ▶ E.g., ImageNet in computer vision.

1. Predicting relative valuations.
   - ▶ Decompose $m_a = \beta_0 + \beta_1 b_{at} + m_a^\perp$.
   - ▶ Estimate $m_a^\perp = \gamma_0 + \gamma_1' x_a + \epsilon_a$ on 80% of the sample.
   - ▶ Evaluate using the R^2 on the remaining 20%.

2. Predicting missing characteristics.
   - ▶ Estimate $c_a = \delta_0 + \delta_1' x_a + \epsilon_a$ on 80% of the sample.
   - ▶ Evaluate using the R^2 on the remaining 20%.

3. Asset similarity in managed portfolios.
   - ▶ Mask the second position of a fund.
   - ▶ Estimate the probability of the identity of the second holding using embeddings/characteristics.

- ▶ Our cross-sectional benchmarks can discriminate between models using a single quarter of data.
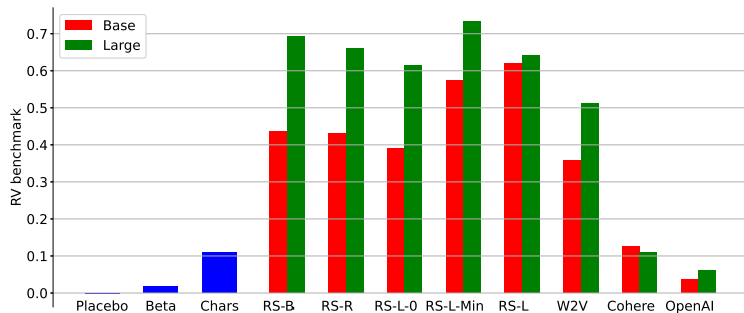
BENCHMARK 1: PREDICTING RELATIVE VALUATIONS
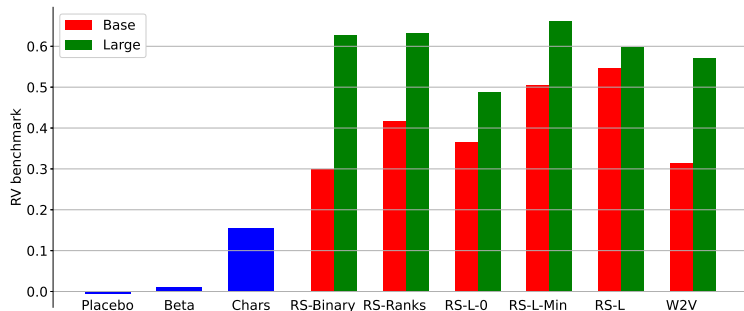
▶ Main takeaways:

 ▶ Holdings-based asset embeddings perform well relative to characteristics.

 ▶ High-dimensional models perform significantly better.
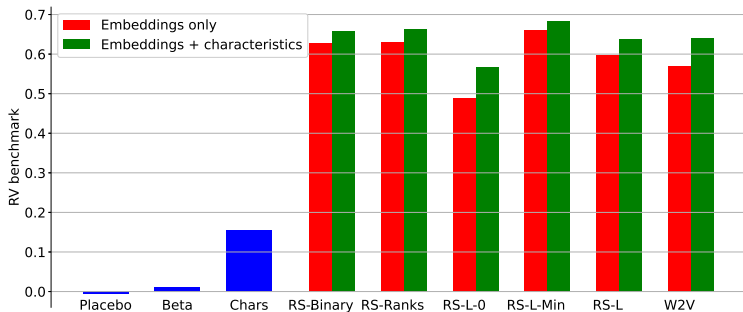
▶ Main takeaway:
  ▶ Adding characteristics to asset embeddings does not improve the benchmark much.

TEXT-BASED EMBEDDINGS

- Main takeaway:
  - Text-based asset embeddings do not perform well.

UNDERSTANDING TEXT-BASED EMBEDDINGS

▶ Using OpenAI's text-based embeddings, we search for the most similar firms (using cosine similarity).

▶ OpenAI's embeddings mix economic and semantic similarity.

Similar Firms as predicted by OpenAI

| Input company	Apple Inc	Citigroup Inc	Walmart Inc
Rank 1	Appian Corp	Citizens Financial Group Inc	Walgreens Boots
Rank 2	Adobe Inc	Goldman Sachs Group Inc	Home Depot Inc
Rank 3	Interdigital Inc	American International Group Inc	Murphy Usa Inc
Rank 4	Microsoft Corp	Comerica Inc	Amazon Com Inc
Rank 5	Gopro Inc	Cigna Corp New	Qurate Retail Inc
Rank 6	Netapp Inc	Capital One Financial Corp	Big Lots Inc
Rank 7	Intel Corp	Caci International Inc	Burlington Stores
Rank 8	Alphabet Inc	Capital City Bank Group	Dollar Tree Inc
Rank 9	Autodesk Inc	C N O Financial Group Inc	Nordstrom Inc
Rank 10	Appfolio Inc	Jpmorgan Chase & Co	Kohls Corp

High-dimensional embeddings
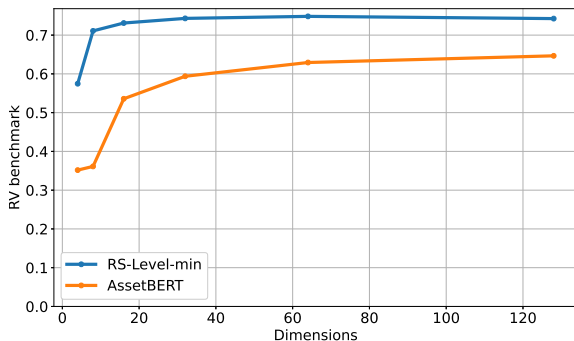
▶ Main takeaways:
  ▶ High-dimensional models perform particularly well.
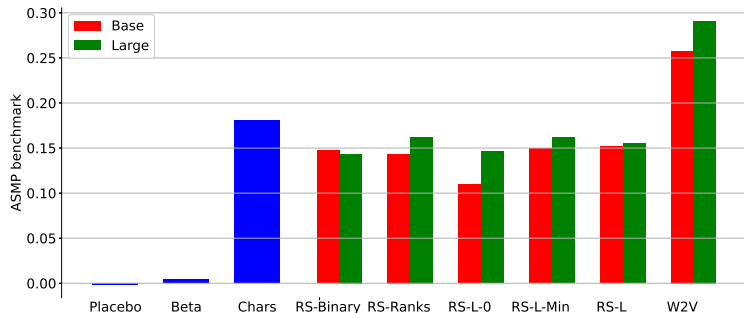 ▶ AssetBERT performs well, but outperformed by the simpler recommender system.

BENCHMARK 2: MISSING CHARACTERISTICS
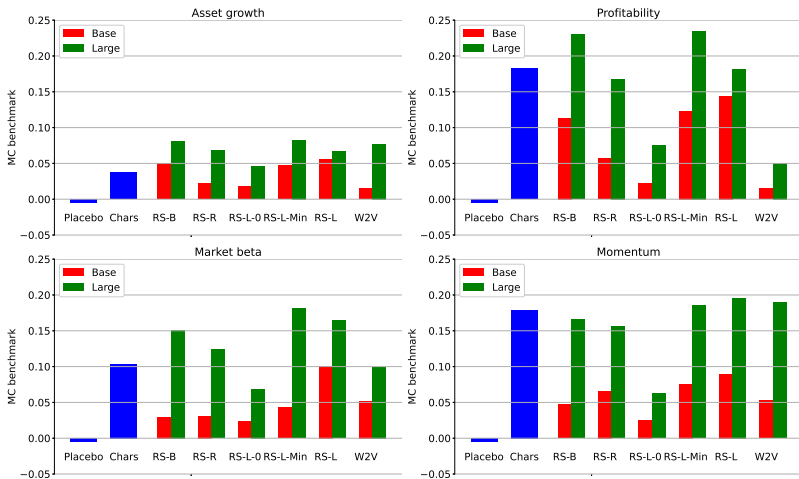
► Main takeaway:
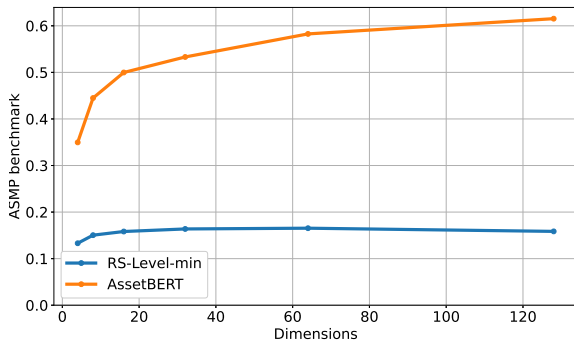 ► Asset embeddings contain useful information to complete the panel of asset characteristics.

BENCHMARK 3: ASSET SIMILARITY

▶ Main takeaway:
  ▶ Word2Vec performs significantly better than recommender systems and observed characteristics.

HIGH-DIMENSIONAL EMBEDDINGS

▶ Main takeaway:
  ▶ AssetBERT performs better than the simpler recommender system and Word2Vec on this benchmark.

GENERATIVE PORTFOLIOS

▶ The AssetBERT model can be used to predict which stocks (right column) fit well in a given portfolio (left column).

	Worst Performing Stocks in 2020 Q1		Similar stocks predicted by AssetBERT	
	Company	Return	Company	Return
1	Occidental Petroleum Corp	-0.70	Metlife Inc	-0.39
2	Marathon Petroleum Corp	-0.60	Cummins Inc	-0.24
3	Discover Financial Services	-0.58	Wells Fargo & Co New	-0.46
4	Eog Resources Inc	-0.57	International Paper Co	-0.31
5	Boeing Co	-0.54	Prudential Financial Inc	-0.44
6	Pioneer Natural Resources Co	-0.53	General Motors Co	-0.42
7	Conocophillips	-0.52	Public Service Enterprise Gp Inc	-0.23
8	American International Group Inc	-0.52	Capital One Financial Corp	-0.51
9	Phillips 66	-0.51	Sysco Corp	-0.46
10	Valero Energy Corp New	-0.51	T Rowe Price Group Inc	-0.19

▶ This model can be improved further using causal language models (i.e., the GPT decoder structure).

INVESTOR SIMILARITY

▶ The simple recommender system already leads to a reasonable clustering of investors.

Fund	Dimensional US Large Cap Value ETF
Rank 1	SA US Value Fund
Rank 2	Dimensional Funds ICVC - International Value Fund
Rank 3	PGIM Quant Solutions Large-Cap Value Fund
Rank 4	UBS (Irl) ETF Plc - Factor MSCI USA Prime Value ESG UCITS ETF
Rank 5	Columbia Multi Manager Value Strategies Fund
Fund	iShares Exponential Technologies Index ETF
Rank 1	Multi Units LU - Lyxor MSCI Disruptive Tech. ESG Filtered
Rank 2	LUX IM - AI & DATA
Rank 3	AtonR Fund (The)
Rank 4	Dux Umbrella FI - Trimming USA Technology
Rank 5	HANetf ICAV - HAN-GINS Innovative Technologies UCITS ET
Fund	Virtus LifeSci Biotech Products ETF
Rank 1	Global X Genomics & Biotechnology ETF
Rank 2	BNY Mellon Global Fds. Plc - Smart Cures Innovation Fund
Rank 3	WisdomTree BioRevolution Fund
Rank 4	JPMorgan Funds - Thematics - Genetic Therapies
Rank 5	JSS Investmentfonds II - Sustainable Eq. - Future Health

INTERPRETING ASSET AND INVESTOR EMBEDDINGS

- ▶ Asset embeddings yield clusters of stocks.

- ▶ We use OpenAI's `GPT-4o` model to summarize the earnings calls of groups of firms and identify
  - ▶ Main common risks.
  - ▶ Main growth opportunities.
  - ▶ ...

- ▶ To avoid generic risks, we can add a group of firms (sampled across industries) as a reference point.

- ▶ The same logic applies to investor embeddings using, e.g., information in fund prospectuses, analyst reports, et cetera.

CONCLUSIONS

▶ Recent advances in AI/ML can be applied to economics and finance via asset embeddings.

▶ We provide a micro foundation for using holdings data.

▶ We adjust methods that have been successful in related areas (e.g., NLP, vision, ...) to economics:
  ▶ Recommender systems, Word2Vec, transformer models.

▶ In progress:
  ▶ Other asset classes (e.g., fixed income, derivatives, . . .).
  ▶ Generate stress scenarios by simulating investor and asset embeddings, combined with an asset demand system (Koijen and Yogo, 2019).